1

Media Generation

1.1 Introduction

Real Time Communication (RTC) bridges the gap between devices and people, enabling instant, interactive exchanges. RTC supports a two-way dialog across *User Equipment* (UE) including smartphones, tablets, laptops, and even IoT devices (Herrero 2021). Video conferencing, voice calls, and instant messaging are just some examples of RTC. Other examples include IoT devices generating real time data that drive applications like predictive maintenance, remote monitoring, smart home automation as well as industrial IoT. This means that RTC involves a range of technologies with versatile applications that enable legacy and emerging architectures.

In this context, this chapter is about media generation, and it specifically focuses on the generation of signals that facilitate RTC. Given that signals originate in the analog domain, their conversion into digital form becomes necessary at the UEs (Herrero 2023). The process of digitization involves two essential mechanisms referred to as sampling and quantization (Haykin 2009). These processes lead to the conversion of analog signals with infinite precision into digital data of finite size, enabling storage and transmission. To achieve even greater compression, these streams can be processed through media codecs (Chu 2003).

The conversion of a signal from the analog realm to the digital domain introduces distortion. This distortion is further exacerbated during transmission through a communication channel affected by network impairments like loss and latency. Quantifying the degree of this distortion involves the utilization of quality scores and other relevant metrics.

Hands-On Real Time Communications: A Practical Guide to RTC Protocols in Non-3GPP, 3GPP, 4G/5G/6G and IoT Networks, First Edition. Rolando Herrero.

Companion website: www.wiley.com/go/herrero/RTCProtocols

1.2 Signals

Real time communication signals, capable of carrying any digital data, are our focus in this chapter, specifically how they carry speech, audio, and video. This is because speech, audio, and video signals are the most common types of signals used in RTC applications. Speech signals are used for voice calling, audio signals are used for music and other audio content, and video signals are used for video conferencing and streaming.

1.2.1 Speech

Audio and speech signals are sound signals, characterized by fluctuations in air pressure as they travel. These pressure variations are described as waves, commonly known as sound waves. Furthermore, speech signals play a pivotal role in human communication and constitute a subset of these audio signals. A key fact is that their distinctive nature allows them to be processed and encoded using mechanisms distinct from those employed for standard audio signals (Jurafsky and Martin 2000).

The human auditory system perceives sounds ranging from 20 to 20000 Hertz (Hz) (or 20 kHz), with heightened sensitivity to events occurring between 250 Hz and 5 kHz. This range is where speech signals predominantly reside. Vowel sounds derive their energy mainly from 250 Hz to 2 kHz, while voiced consonants like *b*, *d*, and *m* are strongest between 250 Hz and 4 kHz. Unvoiced consonants like *f*, *s*, and *t* have varying intensities and occupy the 2 to 8 kHz range.

For good speech comprehension, it is especially important to have good hearing in the range of 125 Hz to 4 kHz, where unvoiced consonants reside. Figure 1.1 shows a normalized speech sequence recorded over 25 milliseconds. Interestingly, it displays a strong correlation within about 8 milliseconds, matching the pitch period of the sequence. More details of the characteristics of speech, including the formal definition of the pitch period, are presented in Section 1.4.1.

1.2.2 Audio

Audio signals cover a wider range than speech signals, reaching the full human hearing range and even including the ultrasonic range. In terms of practicality, audio signals typically transmit in one direction, while speech signals spread in all directions. Unlike speech signals, which have a clear organization with specific patterns in frequency and amplitude, audio signals lack such structure and are less complex, with less data, allowing for better compression.

Similar to speech signals, audio signals exhibit various attributes beyond their frequency range. The amplitude of a speech signal corresponds to its volume or

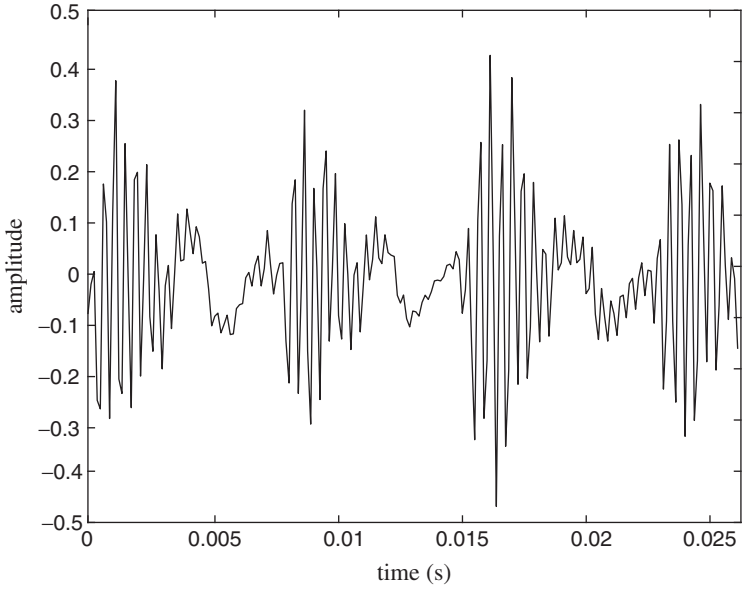


Figure 1.1 Speech Signal.

loudness. This amplitude can vary significantly, influenced by factors like the speaker, surroundings, and the conveyed emotion. Sound pressure amplitude is commonly quantified in units of *pascals* (Pa). In this context, a *micropascal* (μPa) is equivalent to one-millionth of a pascal.

In controlled laboratory settings, the benchmark for the minimum detectable amplitude, known as the threshold of hearing, is approximately 20 μPa for a 1 kHz tone. Some individuals regard 60 μPa as the threshold for experiencing pain, yet this perception is subjective and varies significantly based on individual differences and age. For instance, the sound pressure produced by a jackhammer at a distance of 1 meter reaches a maximum of approximately 2 Pa, while a jet engine at the same distance generates a maximum pressure of 632 Pa. *Sound Pressure Level* (SPL), another parameter of sound field, serves as a metric in acoustic analysis and recording. It is commonly employed to gauge the auditory experience at the listener's location or the positioning of a microphone. The relationship between sound pressure, measured in pascals, and dB SPL is associated with a spectrum spanning from 0.00002 Pa, corresponding to 0 dB SPL, to 200 Pa, equating to 140 dB SPL and even higher (Smith 2010).

Both audio and speech signals exhibit temporal characteristics, such as the zero crossing rate, which refers to the frequency at which the signal crosses the zero amplitude line within a second. Additionally, they manifest time-domain attributes like autocorrelation, illustrating how similar the signal is to itself at

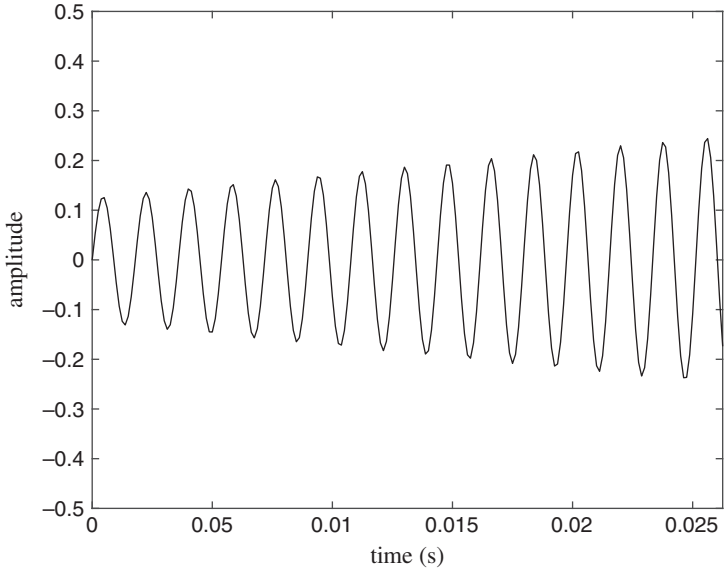


Figure 1.2 Audio Signal.

different time delays. In the frequency domain, these signals possess characteristics including the power spectrum, that show how much energy the signals have at different frequencies. Furthermore, there are *Mel-Frequency Cepstral Coefficients* (MFCCs), which capture the high-frequency content of the signal, and *Linear Predictive Coding* (LPC) coefficients, which model the signal as a linear combination of past samples. Figure 1.1 originally shown in Section 1.2.1 illustrates a speech sequence. In contrast, Figure 1.2 presents a normalized audio sequence of a 2.5 kHz tone with increasing amplitude, showing the differences in temporal and spectral characteristics compared to speech.

1.2.3 Video

The foundation of digital video predominantly originates from technology (now obsolete) introduced during the early stages of black and white television in the 1930s. The receiver component of black and white television featured the *cathode ray tube* (CRT). The CRT utilizes a substantial voltage difference to generate an electric field that propels electrons, giving rise to cathode rays. These rays transport electrons from the cathode to the anode, where, due to their momentum, they collide with a photosensitive screen. The velocity and, therefore, the luminosity on the screen correspond to the voltage applied between the anode and cathode. The CRT also incorporates horizontal and vertical deflection plates, allowing the

path of the rays to be altered and ensuring that every point on the photosensitive screen receives illumination.

The image refresh process happens line by line, starting from the upper left corner and progressing to the lower right. Voltage signals applied to each set of deflection plates have a sawtooth waveform. The y-plate signal's period matches the video frame duration, while the x-plate's period aligns with the line duration. This scanning method is known as progressive scanning. The frame frequency, calculated as the reciprocal of the frame duration, is derived from the *Alternating Current* (AC) frequency to minimize interference from transformers and other power fluctuations. In the United States with 60 Hz AC, the standard frame rate has been traditionally 30 *frames per second* (fps). In regions with 50 Hz AC, like most of the world, the standard has been 25 fps. Note that the first and last lines of each frame, as well as the start and end of each line, are transmitted but not displayed. These hidden regions typically carry synchronization data (Benott 2008).

A different approach from progressive scanning is interlaced scanning. Here, each video frame is divided into two fields, scanned in an alternating manner. The first field scans the odd-numbered lines, while the second scans the even-numbered lines. While progressive scanning delivers one complete frame every $\frac{1}{30}$ seconds (assuming a 30 fps frame rate), interlaced scanning transmits half a frame every $\frac{1}{60}$ seconds, resulting in the same field rate (60 Hz in this example). Keep in mind that, interlaced scanning can provide a perceptual advantage by reducing choppiness in certain situations, like scenes with slow or moderate motion, due to its higher apparent frame rate (twice the field rate). However, it can also introduce artifacts under certain conditions.

Figure 1.3 compares progressive and interlaced scanning. Although both methods process the same number of fields per unit time (operating at $\frac{1}{2T}$ field rate), they achieve this in different ways. Progressive scanning displays each complete frame independently, while interlaced scanning displays half frames

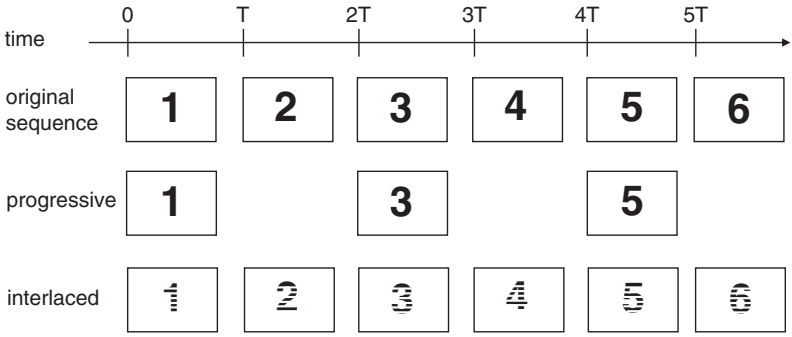


Figure 1.3 Progressive vs Interlaced Scanning.

in an alternating pattern, creating the illusion of a full frame. When viewed in motion under certain conditions, particularly with slow or moderate motion and higher vertical resolutions, interlaced scanning can offer a perception of smoother playback. However, it can also introduce artifacts, especially in fast-moving scenes or at lower resolutions.

Irrespective of the frame rate, all black and white standards utilize a composite analog signal for transmission, incorporating *video, blanking, and synchronization* (VBS) information. This transmission method commonly employs interlaced scanning, where each line's initiation is guided by precisely timed horizontal synchronization pulses embedded between the video signals in the imperceptible portion of each line. The term *blanking* has to do with the pulse responsible for guiding the beam's retrace from the lower right to the upper left corner of the screen.

The black and white television standard in the United States uses a field frequency of 60 Hz, divided into two fields, each containing 262 ½ lines, and a line frequency of 15.75 kHz. Video content uses *Amplitude Modulation* (AM) with a bandwidth of 4.2 MHz, while audio uses *Frequency Modulation* (FM) with a 4.5 MHz carrier frequency.

Unlike the United States standard, the European (and international) black and white television standard operates at a field frequency of 50 Hz. It contains two fields, each composed of 312 ½ lines, resulting in a line frequency of 15.625 kHz. Video content uses AM modulation with a bandwidth of 5 MHz, while audio utilizes FM modulation with a 5.5 MHz carrier frequency.

Since any color can be made by mixing the three primary colors: *red, green, and blue* (RGB), a color TV scheme uses three separate black and white cameras, each with a filter that only lets in one of these primary colors. These cameras capture three separate VBS signals, which can be combined to recreate the original full-color image. However, there is a problem with this approach: old-fashioned black and white TVs cannot display video recorded with this method.

The primary objective is to devise a compatibility scheme that functions bidirectionally–enabling color signals to be displayed on traditional black and white televisions, while also ensuring that monochromatic signals can be exhibited on color televisions.

A linear combination can be applied to the RGB components and transformed into other three components that decompose the information into signals that provide backward compatibility. One such scheme is through a set of linear equations given by

$$
\begin{aligned}
Y &= 0.587G + 0.299R + 0.1145B \\
C_b &= 0.564(B - Y) \\
C_r &= 0.713(R - Y)
\end{aligned}
$$

where R, G, and B are the intensity values of the red, green, and blue components that serve as input to obtain luminance Y, blue chrominance C_b and red chrominance C_r components. An alternative set of equations results from

$$Y = 0.587G + 0.299R + 0.1145B$$
$$U = 0.493(B - Y)$$
$$V = 0.877(R - Y)$$

where the blue and red chrominances are U and V, respectively. Figure 1.4 illustrates how backward compatible signals are generated from the input signals. Each signal involves filtering a specific color of light through optics and then capturing its electrical representation via a regular black and white camera.

The luminance signal conveys the intensity of the color image, allowing black and white TVs to display a grayscale version. For color TVs, the chrominance signals carry the color information needed to render the complete image. Figure 1.5 depicts the composite signal, which combines these elements. Figures 1.6 to 1.8 further break down the Y, U, and V components, respectively, that make up the chrominance signal. Interestingly, human perception is less sensitive to color compared to luminance. This allows less bandwidth to be allocated for transmitting chrominance details, ensuring compatibility with older standards without impacting the critical black and white components.

The VBS signal containing chrominance information is called *Color VBS* (CVBS). Different ways of modulating chrominance signals create various color representation schemes, leading to different video transmission standards. For example, the United States relied on the *National Television Standard Committee* (NTSC) standard, while Europe utilized both *Phase Alternating Line* (PAL) and

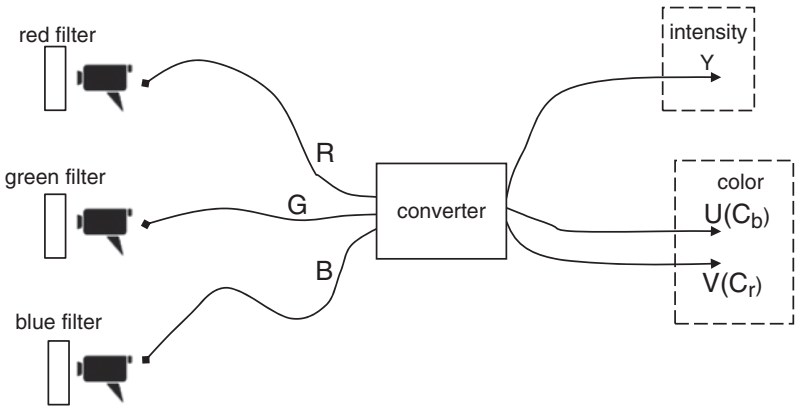


Figure 1.4 RGB to YUV Conversion.

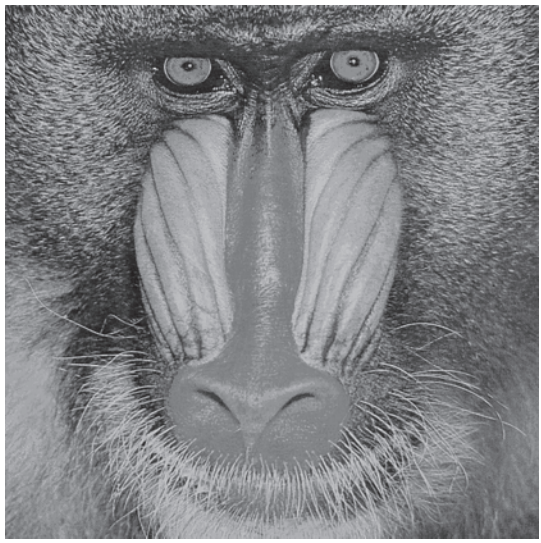


Figure 1.5 Composite Image.

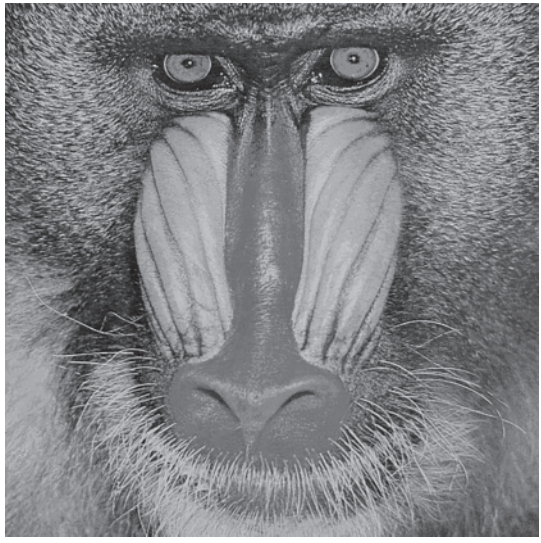


Figure 1.6 Y Component.

Sequential Color with Memory (SECAM), which were once dominant players in analog television broadcasting. Although analog TV is no longer used, these seemingly outdated technologies laid the foundation for many aspects of modern digital video. Details of digital video coding are introduced later in this chapter in Section 1.4.3.

Figure 1.7 U Component.

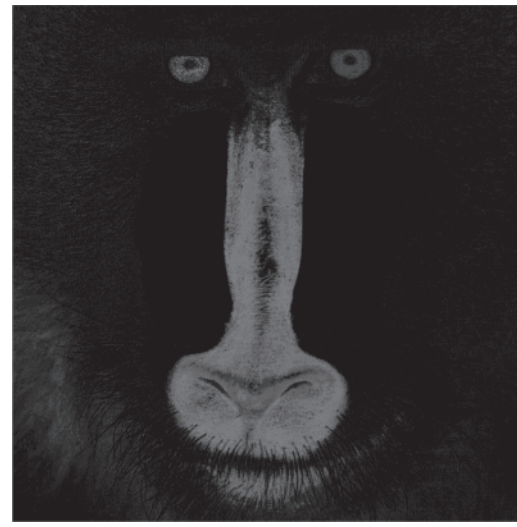


Figure 1.8 V Component.

1.3 Sampling and Quantization

As outlined in Section 1.1, information produced by natural sources such as speech, audio, and images exists in the analog domain. In this domain, corresponding signals often include a significant amount of redundant data. This redundant information can be eliminated without appreciably affecting human

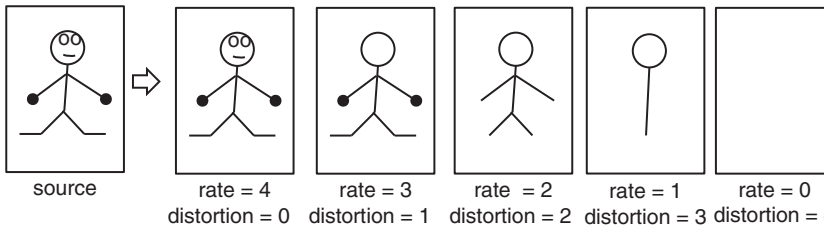


Figure 1.9 Rate–Distortion Example.

perception, a procedure referred to as data compression. It is worth noting that, in contrast to the analog domain, a digital domain exists where analog signals are translated into sequences derived from a countable set of numbers. These sequences can be stored and processed by signal processors and computers. The process of removing information in this context is termed source encoding.

Data compression can be divided into two primary categories. First, there is lossless compression, which involves eliminating redundancy in a reversible manner, ensuring that the reconstructed message remains identical to the original. Second, there is lossy compression, which entails purposefully removing information in a controlled manner. This removal of non-essential data is permanent, rendering the process irreversible. In this scenario, although the reconstructed message differs from the original, the objective is to minimize this divergence to the point of imperceptibility to the recipient.

The compression rate is defined as the ratio of the information present in the compressed message to that in the original message. Generally, lossy compression exhibits a significantly higher compression rate compared to lossless compression.

In the realm of lossy compression, a specific compression rate corresponds to a specific degree of distortion, giving rise to a trade-off known as rate–distortion. An illustration of this concept can be observed in Figure 1.9, where a source image undergoes compression at various rates, leading to distinct levels of distortion. Both rate and distortion are quantified on a unitless scale ranging from 0 to 4. Note that as distortion diminishes, transmission rates tend to increase; conversely, higher distortion levels result in lower transmission rates (Proakis and Manolakis 2006; Haykin 2009).

As indicated in Section 1.1, sampling and quantization are two key mechanisms that must be employed to convert signals from the analog to the digital domain. The subsequent subsections focus on these two key processes.

1.3.1 Sampling

Sampling transforms an analog signal, which is defined continuously across time, into a sequence of discrete samples taken periodically at a fixed interval known

as the sampling period, measured in units of time. The reciprocal of the sampling period is the sampling rate, measured in units of *samples per second* (sps). If the sampling rate is too slow, the resulting set of samples might not accurately represent the original analog signal. Conversely, if the rate is excessively high, the set could become unwieldy for efficient processing and modulation for transmission through a channel. Table 1.1 shows how, in audio and speech sampling scenarios, each sampling rate has a specific type that identifies it.

Generally, when an energy signal $g(t)$ (as shown in Figure 1.10) is defined at every instant of time and sampled at a rate of $F_s = \frac{1}{T_s}$ (where T_s denotes the sampling period), the resulting sampled signal resembles Figure 1.11. Mathematically, sampling takes $g(t)$ as input and generates an infinite sequence of samples spaced T_s seconds apart, forming the sequence $\{g(nT_s)\}$. This process, known as ideal or instantaneous sampling, captures the signal's value at specific, infinitesimal

Table 1.1 Sampling Rate Types.

Type	Sampling Rate (sps)
Narrowband (NB)	8000
Wideband (WB)	16000
Super Wideband (SWB)	32000
Fullband (FB)	48000

Figure 1.10 Signal $g(t)$.

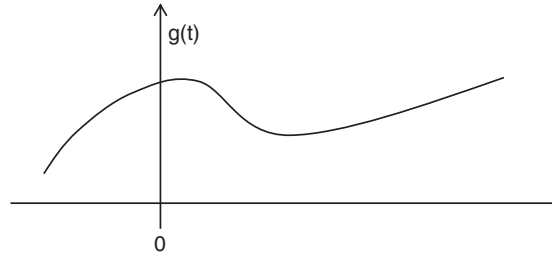
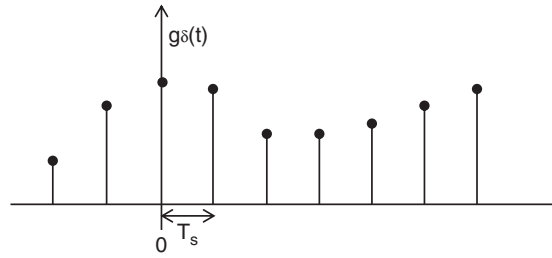


Figure 1.11 Sampled $g(t)$.



instants of time. Note that under ideal sampling, the effect is akin to multiplying the analog signal $g(t)$ by a train of time pulses. Specifically,

$$\begin{aligned} g_\delta(t) &= \sum_{n=-\infty}^{\infty} g(nT_s)\delta(t - nT_s) \\ &= g(0)\delta(t) + \\ &\quad g(T_s)\delta(t - T_s) + \\ &\quad g(2T_s)\delta(t - 2T_s) + \cdots \end{aligned} \tag{1.1}$$

where $g_\delta(t)$ is the ideal sampled signal and $\delta(t - nT_s)$ is a delta function positioned at time $t = nT_s$.

In the frequency domain $g_\delta(t)$ becomes $G_\delta(f)$ and it is given by

$$\begin{aligned} G_\delta(f) &\rightleftharpoons F_s \sum_{m=-\infty}^{\infty} G(f - mF_s) \\ &= F_s G(f) + \\ &\quad F_s G(f - F_s) + F_s G(f + F_s) + \\ &\quad F_s G(f - 2F_s) + F_s G(f + 2F_s) + \cdots \end{aligned} \tag{1.2}$$

where the spectrum is an infinite sequence of shifted versions of the analog signal frequency representation. Whether overlap occurs among these versions hinges upon the sampling rate F_s. To illustrate, consider a scenario where $g(t)$ is band-limited (represented in Figure 1.12), containing no frequency components beyond W Hz. If sampled at a rate of $F_s = 2W$, the resulting spectrum (as shown in Figure 1.13) includes components without any overlap.

Mathematically, from Equation (1.2)

$$G_\delta(f) = F_s G(f) + F_s \sum_{m=-\infty, m\neq 0}^{\infty} G(f - mF_s)$$

where if (1) $G(f) = 0$ for $|f| \geq W$ and (2) $F_s = 2W$, then

$$G(f) = \frac{1}{2W} G_\delta(f) \tag{1.3}$$

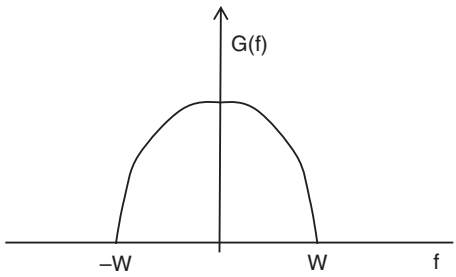


Figure 1.12 Band-limited Signal.

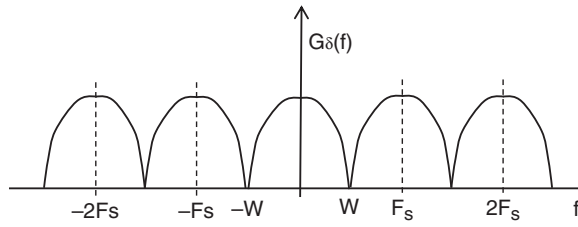


Figure 1.13 Spectrum of Sampled Signal.

if $|f| < W$. In this scenario, the samples $g(\frac{n}{2W})$ of the analog signal taken every $\frac{1}{2W}$ seconds, with n being an integer, contain all the information in $g(t)$.

The inverse operation, that is, recovering the analog signal $g(t)$ out of the samples $g(nT_s)$ is given by

$$\begin{aligned} g(t) &= \sum_{n=-\infty}^{\infty} g\left(nT_s\right)\ \mathrm{sinc}\left(2Wt - n\right) \\ &= \sum_{n=-\infty}^{\infty} g\left(nT_s\right)\ \mathrm{sinc}\left(\frac{t}{T_s} - n\right) \\ &= g(0)\,\mathrm{sinc}\left(\frac{t}{T_s}\right) + \\ &\quad g(T_s)\,\mathrm{sinc}\left(\frac{t}{T_s} - 1\right) + \\ &\quad g(2T_s)\,\mathrm{sinc}\left(\frac{t}{T_s} - 2\right) + \cdots \\ &= g_\delta(t) * \mathrm{sinc}\left(2Wt\right) \end{aligned} \tag{1.4}$$

where delayed versions of the sinc function are added together to interpolate $g(t)$ for any value of t. Note that the sinc function is defined as $\mathrm{sinc}\,(x) = \frac{\sin(x)}{x}$. This convolution in the time domain is analogous to multiplication by a low-pass (LP) filter in the frequency domain, referred to as a reconstruction filter. As a result, to recover the analog signal from its sampled version, $g_\delta(t)$, it is only necessary to process it through this reconstruction filter.

Generally, an energy signal lacking frequency components above W Hz can be accurately described using samples taken periodically every $T_s = \frac{1}{2W}$ seconds. The sampling rate $F_s = \frac{1}{T_s} = 2W$ is known as the Nyquist Rate, while the sampling period $T_s = \frac{1}{F_s} = \frac{1}{T_s}$ is referred to as the Nyquist Interval.

Undersampling occurs when a signal is sampled at a rate below the Nyquist Rate, indicated by $F_s < 2W$. This condition leads to aliasing, a phenomenon where successive shifted replicas of the analog signal's frequency representation in Equation 1.2 overlap. To illustrate, consider the frequency representation of the signal shown in Figure 1.12. When undersampled, as shown in Figure 1.14, the signal exhibits the characteristic distortions of aliasing.

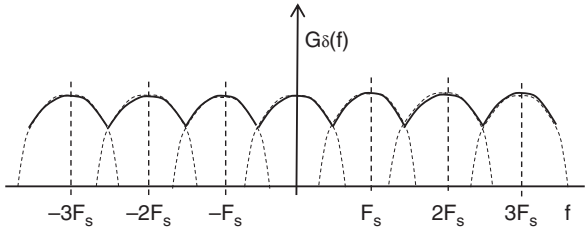


Figure 1.14 Aliasing.

To avoid aliasing, an LP filter, known as an anti-aliasing filter, can be used on the analog signal $g(t)$. This filter suppresses high-frequency components that could otherwise trigger undersampling issues. However, since this filter is not inherently perfect and possesses a transition band, it is important to make slight adjustments to the sampling rate. By ensuring a rate marginally higher than the Nyquist Rate, the absence of spectral overlap can be guaranteed.

When the signal shown in Figure 1.12 is sampled at a rate surpassing the Nyquist Rate, the resulting spectrum of the instantaneous sampled version, shown in Figure 1.15, is achieved.

Reconstruction and anti-aliasing filters typically share analogous characteristics, both possessing a transition band that extends from W to $F_s - W$. Moreover, as the sampling frequency increases, the spectral gap between repetitions of the analog signal's spectrum within the sampled signal's spectrum widens. This phenomenon alleviates constraints on the transition band's width in these filters, illustrated in Figure 1.16. The ultimate extent of this band is dictated by $F_s - 2W$.

Illustrated in Figure 1.17 is an illustrative sampling scenario, examined from the viewpoints of both time and frequency domains. To be precise, the signal designated for sampling at the transmitter undergoes multiplication and convolution by a series of pulses in both the time and frequency realms, respectively. This resultant signal is then transmitted through the communication channel and subsequently reconstructed at the receiver by means of frequency multiplication utilizing an LP filter. Correspondingly, in the temporal domain, this equates to convolution through the time depiction of the LP filter, which manifests as a sinc function.

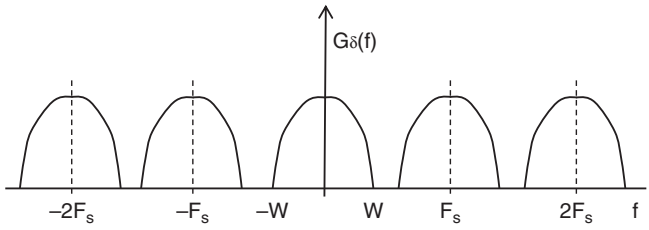


Figure 1.15 Spectrum of Instantaneous Sampled Signal.

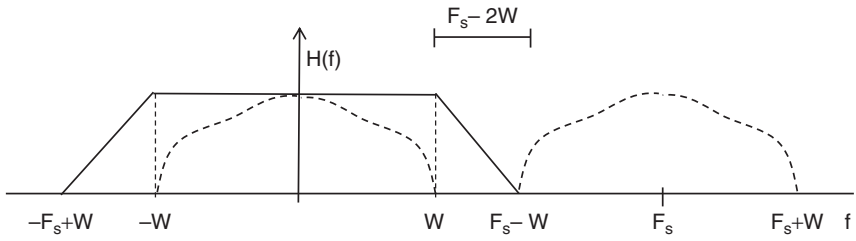


Figure 1.16 Reconstruction Filter.

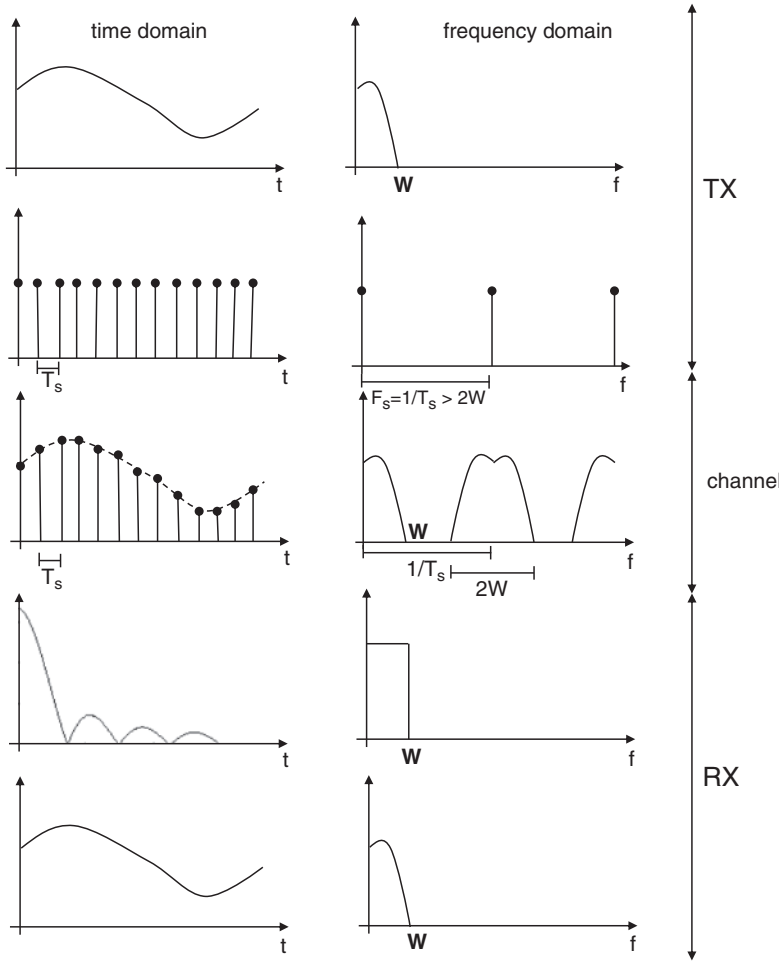


Figure 1.17 Time and Frequency Domain for Sampling.

Figure 1.17 depicts a demonstrative sampling scenario, examined from both time and frequency domain perspectives. Specifically, the signal slated for sampling at the transmitter undergoes multiplication by a series of pulses in the time domain, corresponding to convolution in the frequency realm. The resulting signal traverses the communication channel, and upon reaching the receiver, it is reconstructed via frequency-domain multiplication using an LP filter. In the time domain, this translates to convolution with the time representation of the LP filter, which takes the form of a sinc function (Proakis and Manolakis 2006).

1.3.2 Quantization

While converting analog signals into digital formats, sampling captures a discrete collection of time-domain samples. However, each sample inherently possesses an infinite range of amplitude levels, making their representation using a limited set of digital values impractical. To address this challenge, quantization is employed alongside sampling. This process converts the potential analog sample amplitudes into a finite set of predetermined values, enabling their representation in the digital realm. As expected, quantization inevitably introduces distortion due to its mapping of an infinite range onto a finite set. If the number of available mapped values is insufficient, this distortion can negatively impact human perception.

Accordingly, when the function $m(t)$ is sampled at a frequency of $F_s = \frac{1}{T_s}$, each sample $m(nT_s)$, undergoes transformation into a distinct amplitude, denoted as $v(nT_s)$, selected from a predetermined set of values. Note that quantization typically operates in a memoryless, per-sample manner, guaranteeing that the quantization of one sample remains independent of quantization decisions made for previous samples.

In terms of amplitudes, if an analog sample has an amplitude m, then through quantization, this amplitude is mapped into a number k if m is within a specific range or partition cell, shown in Figure 1.18 and given mathematically by

$$P_k : \{m_k < m \leq m_{k+1}\}, \qquad k = 1, 2, \cdots, L$$

where L is the total number of levels or possible discrete amplitudes in the quantization set and the discrete amplitudes m_k with $k = 1, 2, \cdots, L$ are called decision thresholds. Upon the need to reconstruct the quantized sample as a specific value v, the numerical index k undergoes a conversion process to yield a reconstruction level denoted as v_k. This reconstruction level involves the entire spectrum of conceivable analog amplitudes present within a designated partition referred to as P_k.

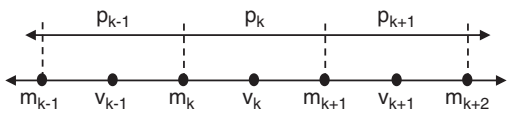


Figure 1.18 Partition Cell.

The interval between two successive levels, represented as $\Delta = v_k - v_{k-1}$, is known as the step size.

In a broad sense, the conversion process from an analog sample to the reconstructed value v is expressed as $v = g(m)$, where the function $g(.)$ is referred to as the quantizer characteristic. Depending on whether it possesses an even or an odd count of levels, the quantizer can take the form of either a midtread or a midrise configuration.

In the process of quantization at the transmitter stage, the analog sample's magnitude is associated with an index denoted as k, which can be transformed into a binary representation, comprising a sequence of bits. This binary sequence is subsequently modulated and transmitted through the communication channel. At the receiver's end, the transmitted signal undergoes demodulation, resulting in a sequence of bits that correspond to the index k. This index is then utilized to reconstruct the original sample's magnitude using the quantizer's characteristics.

Quantization noise, also shown in Figure 1.19 for the case of a ramp signal, is characterized as the disparity between the input signal $m(t)$ and the reconstructed quantized signal $v(t)$. Treating both $m(t)$ and $v(t)$ as random processes, their respective sampling at any given time point yields two *Random Variable*s (RVs) denoted as M and V. The quantization noise is then represented as a third random variable, termed as Q, calculated as the difference between M and V. Assuming a uniform midrise quantizer where both M and Q have zero mean and where the distribution of the amplitude of M is in the range $(-m_{\max}, m_{\max})$ then the step size of the

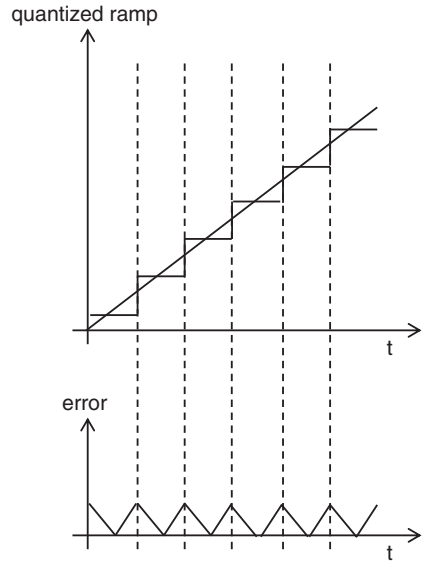


Figure 1.19 Midtrise Noise.

quantizer is given by

$$\begin{aligned}\Delta &= \frac{2m_{\max}}{L} \\ &= \frac{2m_{\max}}{2^n}\end{aligned} \tag{1.5}$$

where n is the number of bits per sample such that $L = 2^n$ is the even total number of reconstruction levels. The quantization error Q can be thought of as a uniformly distributed RV where an instance q is in the interval $-\frac{\Delta}{2} \leq q \leq \frac{\Delta}{2}$. The *Probability Density Function* (PDF) of Q is therefore given by

$$f_Q(q) = \begin{cases} \frac{1}{\Delta} & -\frac{\Delta}{2} < q \leq \frac{\Delta}{2} \\ 0 & \text{otherwise} \end{cases}.$$

The variance of Q obtained from σ_Q^2 and given by

$$\begin{aligned}\sigma_Q^2 &= E\left[Q^2\right] \\ &= \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} q^2 f_Q(q) dq \\ &= \frac{\Delta^2}{12}\end{aligned} \tag{1.6}$$

is a good indication of the quantization error power, which should be as small as possible to minimize distortion. Replacing Equation 1.6 in Equation 1.5, the power of the quantization error results

$$\sigma_Q^2 = \frac{1}{3} m_{\max}^2 2^{-2n}$$

and assuming that the power of the input analog signal $m(t)$ is P, then the output *Signal-To-Noise Ratio* (SNR) of the quantizer is given by

$$\begin{aligned}\text{SNR}_O &= 10 \log_{10}\left(\frac{P}{\sigma_Q^2}\right) \\ &= 10 \log_{10}\left[\left(\frac{3P}{m_{\max}^2}\right) 2^{2n}\right]\end{aligned} \tag{1.7}$$

such that SNR increases exponentially with the number of bits per sample n. Unfortunately, increasing n has the undesirable consequence of increasing both the transmission rate and the bandwidth requirements, as more bits are transmitted over the channel.

Example 1.1 Consider a speech signal with a bandwidth of 3000 Hz that is subjected to a 256-level quantizer. What is the transmission rate of this signal when it is digitalized ?

Solution:
The sampling frequency is $F_s = 2 \times 3000 = 6000$ sps. With $L = 256$ level, the number of bits per sample is $n = \log_2 L = 8$ bits per sample. The transmission rate then becomes $R = F_s \times n = 6000 \times 8 = 48$ Kbps.

1.4 Codecs

Codecs depend on particular signal encoding methods that pertain to the specific type of media being considered (Proakis and Manolakis 2006).

1.4.1 Speech Coding

Based on the source's characteristics, further compression can be applied within the context of source encoding to reduce the overall transmission rate. Specifically, when speech is the source, a speech coder or codec can be utilized to convert speech samples into a bitstream using various techniques. These techniques range from simple input–output mapping to more advanced speech models that extract distinct parameters.

Figure 1.20 illustrates a digital speech communication system in which the transmitter relies on a source encoder responsible for anti-aliasing, sampling, and quantization/symbol encoding. This process results in a bitstream that is subsequently processed by the codec encoder. The codec encoder undergoes channel encoding and modulation before transmission over the channel.

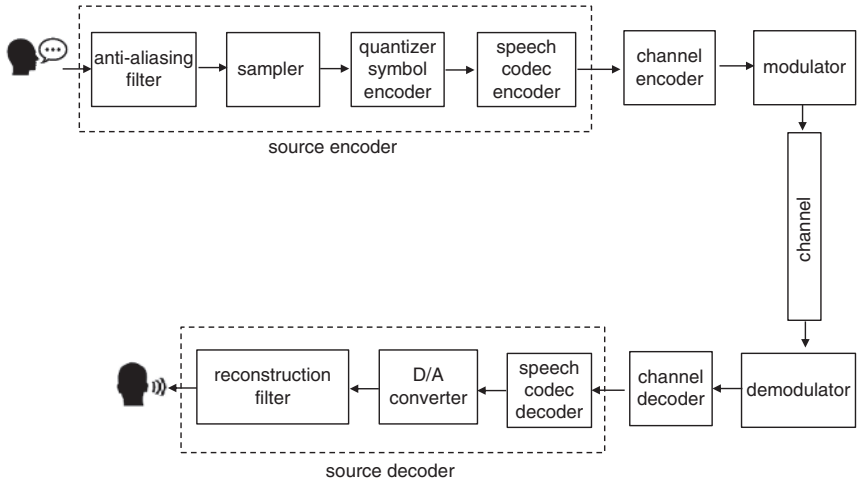


Figure 1.20 Digital Speech Communication System.

Conversely, the receiver performs the inverse operation. After demodulating the modulated wave and decoding the channel, the signal is directed to the codec decoder. The codec decoder restores the signal, which, when subjected to the reconstruction filter, yields the synthetic speech signal. In particular, due to the concentration of speech energy between 300 and 3400 Hz, the speech signal is initially subjected to LP filtering through an anti-aliasing filter with a cutoff frequency of 4000 Hz. To comply with the sampling theorem, a common sampling rate of 8000 sps is often employed. In terms of speech coding, this sampling rate is known as the *narrowband* (NB) sampling rate.

The quantizer used is a midrise uniform scheme with a 16-bit resolution, providing 65536 levels and generally ensuring satisfactory quality for most speech signals. This configuration leads to a bitrate of 8000 sps × 16 bits/sample = 128 Kbps at the input of the codec encoder. Since transmission rates are measured in *bits per second* (bps) units, 1 Kbps is equivalent to 1000 bps. The primary objective of the codec encoder is to compress the raw speech input into an output bitstream lower than 128 Kbps. This scheme is traditionally known as a *Pulse Code Modulation* (PCM) scheme.

Depending on the compression mechanisms employed, a large number of output transmission rates can be achieved. For instance, codecs like ITU-T Recommendation G.711 and AMR (described later in this chapter) exhibit output transmission rates of 64 Kbps and 4.75 Kbps, respectively. This variability is shown in Figure 1.21, where the placement of a codec encoder and decoder in succession maintains an output bitstream of 128 Kbps at the decoder stage.

Within the realm of modern RTC reliant on packet-switched communication networks, the codec encoder plays a key role in buffering raw speech and subsequently compressing it into smaller output frames. To elaborate, the speech codec encoder takes in an input bitstream and generates an output frame. This resultant output frame is then subjected to channel encoding, involving the addition of headers to create packets which are subsequently modulated and sent over the communication channel.

To mitigate the impact of channel encoding overhead on transmission rate, the codec encoder must maintain a relatively larger frame size. However, it is important to have a balance, as an excessively large frame size could introduce latency, negatively affecting the overall user experience. Consequently, there exists a trade-off between frame size and communication latency, ultimately determining

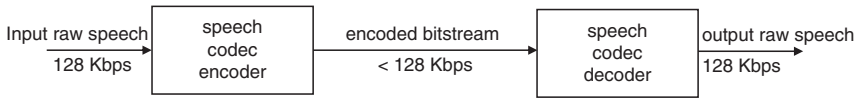


Figure 1.21 Encoder / Decoder.

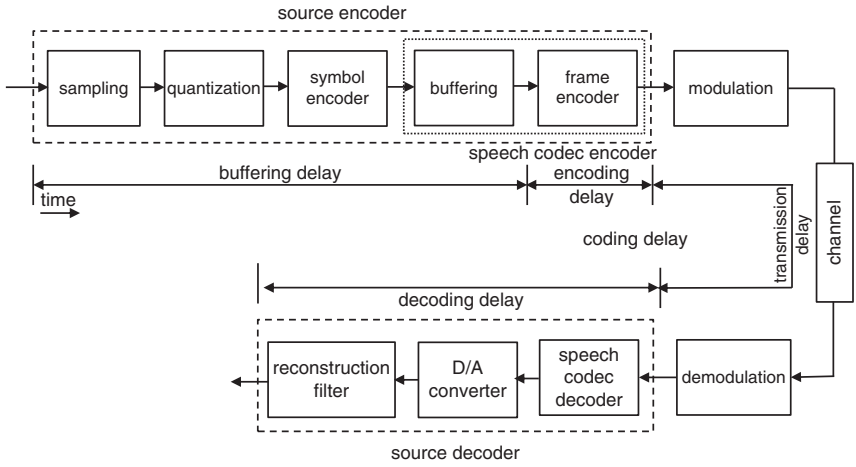


Figure 1.22 Coding Delay.

the optimal frame length. This ideal frame length aims to simultaneously achieve the lowest possible latency and the lowest transmission rate.

The primary objective of a codec is to achieve a substantial reduction in transmission rate through speech compression, while also ensuring quality, resilience against channel errors, commendable performance across diverse languages and ideally compatibility with audio signals like music and tones. From a technical point of view, codecs aim to maintain low computational and memory demands as well as relatively minimal encoding and decoding delays.

The overall delay, as illustrated in Figure 1.22, consists of four distinct components: buffering delay, encoding delay, transmission delay, and decoding delay. Within a speech codec encoder, a buffering element accumulates a sequence of samples into an input frame, which is then compressed by a frame encoding element to generate an output frame. These two elements operate sequentially to achieve speech compression.

The buffering delay, calculated as the time difference between speech entering the source encoder and the input frame buffer reaching the frame encoder, is primarily influenced by the sampling rate and input frame length. For example, with speech sampled at 8000 sps and a frame length of 160 samples, the buffering delay is $\frac{160}{8000} = 20$ ms. Common buffering delays typically range within 10, 20, and 30 milliseconds.

The encoding delay, which is the time required to compress an input frame into an output frame, depends on both the complexity of the compression algorithm and the processing speed of the system. In a pipelined setup, where frames are

processed sequentially, encoding must be faster than buffering to prevent backlogs and ensure smooth operation.

Transmission delay refers to the time it takes for a signal to travel from modulation into the channel to demodulation upon exiting the channel. This delay can exhibit two distinct modes, each associated with a different switching mechanism. Constant transmission aligns with circuit switching and occurs when the transmission delay precisely matches the buffering delay. Similarly, bursty transmission corresponds to packet switching and happens when the transmission delay is shorter than the buffering delay.

Figure 1.23 shows these two transmission modes. Further details about circuit and packet switching mechanisms are discussed in Section 1.5.1. Note that the decoding delay is the time required for the demodulated signal to pass through the source decoder to produce the final synthetic speech. Speech codecs are classified based on the transmission rate of the compressed speech they produce, falling into the following categories:

- *High Bit Rate* (HBR) codecs: Exceed 15 Kbps.
- *Medium Bit Rate* (MBR) codecs: Operate between 5 and 15 Kbps.
- *Low Bit Rate* (LBR) codecs: Range from 2 to 5 Kbps.
- *Very Low Bit Rate* (VLBR) codecs: Fall below 2 Kbps.

Waveform codecs, as opposed to other codec types based on their coding mechanism, preserve the shape of the original speech waveform in the synthesized output. This waveform fidelity grants them adaptability to diverse input signal types, with SNR serving as a reliable quality metric. Due to their inherent characteristics, waveform codecs typically align with the HBR classification (Chu 2003). Parametric codecs, in contrast to waveform codecs, operate by extracting key representative

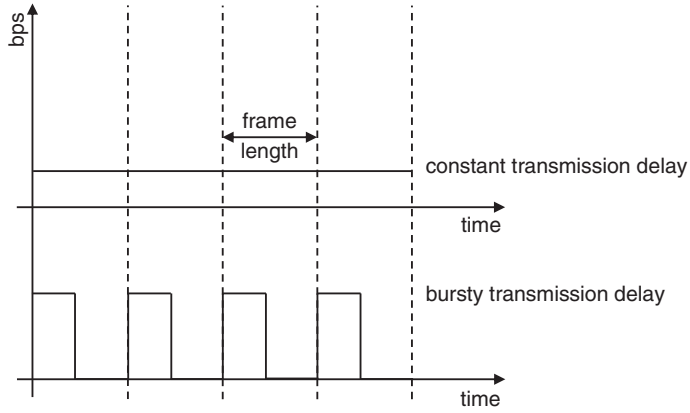


Figure 1.23 Transmission Delay; Constant vs Burst.

parameters from speech sequences. These parameters, rather than the complete waveform, are encoded and transmitted, leading to lower bit rates. However, this approach typically results in a loss of waveform fidelity, making SNR an unsuitable quality metric for parametric codecs. Due to their focus on parameter extraction, they often fall into the LBR or VLBR categories. Hybrid codecs, bridging the gap between waveform and parametric approaches, offer a balanced solution for speech compression. They extract parameters from speech sequences, similar to parametric codecs, but integrate enhancements to preserve the shape of the original waveform in the synthesized output. This combination of techniques often positions hybrid codecs within the MBR category. Waveform and parametric encoders diverge in their treatment of raw frames. Waveform encoders process each sample within a frame individually, while parametric encoders exhibit a different approach, generating compressed output frames distinct from their original inputs.

Understanding and extracting parameters from speech relies heavily on recognizing different signal types. Voiced signals result from the periodic vibration of vocal cords, creating a succession of sounds marked by a characteristic periodicity. This translates to a periodic waveform with a well-defined pitch period and its inverse, the pitch frequency. This frequency typically ranges from 50 to 500 Hz, depending on the speaker's gender. In contrast, unvoiced signals lack any such periodicity and display a noise-like character.

Figure 1.24 contrasts unvoiced and voiced signals within a speech sequence of the phrase *Thank You*. The syllable *Th* exemplifies unvoiced signals with its random, noise-like waveform, while the syllable *U* exhibits the periodic nature of voiced signals. While speech is generally considered non-stationary due to its inherent variability, it can be approximated as stationary over short intervals for analytical purposes. These intervals, characterized by relatively stable signal properties, align with the frame sizes employed by speech codecs. By processing speech on a per-frame basis, codecs manage its non-stationary nature. Frame sizes are further optimized to minimize latency and overhead associated with transport headers.

In speech processing with an NB sampling rate, frame lengths of 10, 20, and 30 milliseconds are commonly employed. These lengths correspond to frame sizes of 80, 160, and 240 samples, respectively.

The speech model foundational to parametric codecs is illustrated in Figure 1.25. This model conceptualizes speech production as a process initiated by airflow from the lungs, which then excites the vocal tract. Airflow is represented as a white noise source, while the vocal tract is modeled as an adaptive, time-varying filter. The filter parameters mirror the acoustic properties of various vocal tract structures, including the nasal cavity, trachea, nostrils, and others. Techniques like LPC are employed to accurately determine these filter parameters.

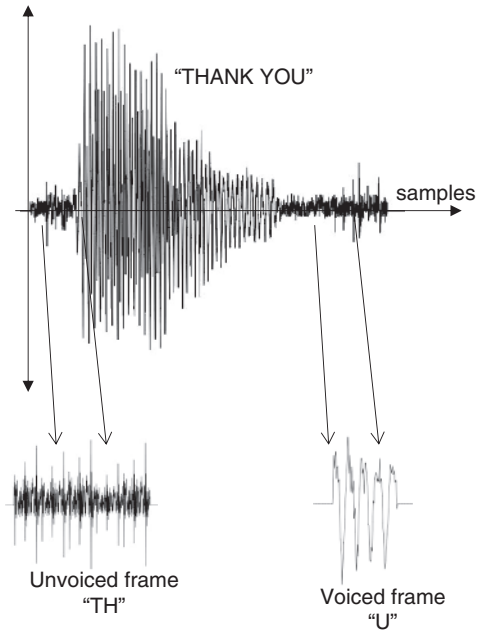


Figure 1.24 The Phrase "Thank You".

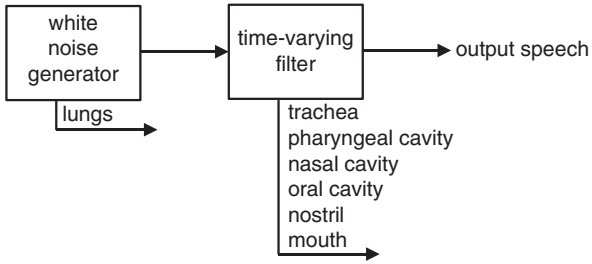


Figure 1.25 Speech Production Model.

1.4.1.1 Waveform Codecs

As indicating in the preceding section, codecs that maintain the original form of the source waveform are referred to as waveform codecs. One of the mechanisms used by these codecs is known as *compression-expansion* (companding). Companding is applied to PCM streams to adjust the speech's dynamic range while preserving the fundamental signal structure.

In traditional PCM systems, analog domain companding takes place prior to numerical conversion, as illustrated in Figure 1.26. This analog companding process can be digitally simulated on a per-frame basis by converting 16-bit samples into 8-bit companded samples using dedicated conversion tables. The underlying concept involves the symbol encoder transforming each of the 65536 sampled and quantized symbols into signed 16-bit numerical representations, accurately

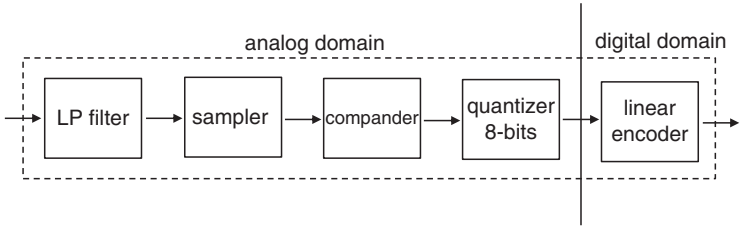


Figure 1.26 μ-Law Analog Companding.

reflecting the analog amplitude of each sample. Note that the maximum positive and negative deviations of an analog sample are mapped to values 32768 and −32767, respectively.

The continuity of analog speech waveforms leads to minimal value variation between consecutive samples, opening a pathway for efficient transmission rate reduction through differential encoding. This technique capitalizes on the narrower range of differences between adjacent samples compared to the range of the individual samples themselves. By operating within a reduced input dynamic range, differential encoding achieves a higher output resolution while maintaining the same number of mapping levels, conserving transmission bandwidth. This technique known as *Differential Pulse Code Modulation* (DPCM) supports this quantization principle and enables efficient speech transmission. A first approximation of DPCM consists of calculating the differential or error signal $e(t)$ as

$$\begin{aligned} e(n) &= x(n) - x_p(n) \\ &= x(n) - x(n-1) \end{aligned} \tag{1.8}$$

where $x(n)$ and $x_p = x(n-1)$ are the current and the predicted 16-bit input samples, respectively. The error signal is then quantized as $\hat{e}(n)$ in accordance with a number of possible levels and transmitted over the channel. At the receiver, the quantized error is used to generate the output sample $\hat{x}(n)$ given by

$$\begin{aligned} \hat{x}(n) &= \hat{e}(n) + \hat{x}_p(n) \\ &= \hat{e}(n) + \hat{x}(n-1) \end{aligned} \tag{1.9}$$

where, for the same sample number, $\hat{x}_p(n)$ is different than $x(n)_p$ at the transmitter. The result of this scheme is that for a 2-level quantizer, the resulting $\hat{x}(n)$ is different than $x(n)$.

A solution to this problem consists on a scheme where the quantized error $\hat{e}(n)$ is related to the quantized prediction error $\hat{x}_p(n)$ as

$$\begin{aligned} \hat{x}(n) &= \hat{e}(n) + \hat{x}_p(n) \\ &= \hat{e}(n) + \hat{x}(n-1) \end{aligned} \tag{1.10}$$

on both, encoder and decoder. The figure also shows, that even if the same 2-level is used, $\hat{x}(n)$ also follows the original input 16-bit signal $x(n)$.

To compare the performance of PCM and DPCM, let us examine the triangular signal illustrated in Figure 1.27. When this signal undergoes a 16-level PCM scheme, each 16-bit sample is converted into a 4-bit PCM sample. The reconstruction of this signal, however, introduces errors, as shown alongside the output signal in Figure 1.28. This error results in an SNR of 24 dB.

Likewise, when subjecting the identical signal to a 16-level DPCM scheme, where an error is computed for each 16-bit sample and subsequently quantized into a 4-bit DPCM sample, the resulting reconstruction error is evident. This error is displayed alongside the output signal in Figure 1.29, leading to a corresponding SNR of approximately 49.42 dB. In particular, the SNR disparity between DPCM and PCM amounts to roughly 25 dB, representing an improvement of approximately 340 times in favor of DPCM. This enhancement in performance is achieved by diminishing the dynamic range of the quantized signal while retaining the same number of quantization levels, thereby upholding the transmission rate.

Applying a 16-level DPCM scheme to the same triangular signal yields distinct results. In this case, an error is calculated for each 16-bit sample and quantized into a 4-bit DPCM sample. As shown alongside the output signal in Figure 1.29, the reconstruction error is reduced, leading to a significantly higher SNR of approximately 49.42 dB. This disparity of 25 dB between DPCM and PCM, translating to a 340-fold improvement in SNR, is about the effectiveness of differential encoding.

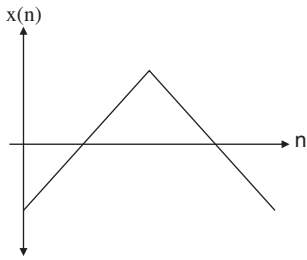


Figure 1.27 Input Signal.

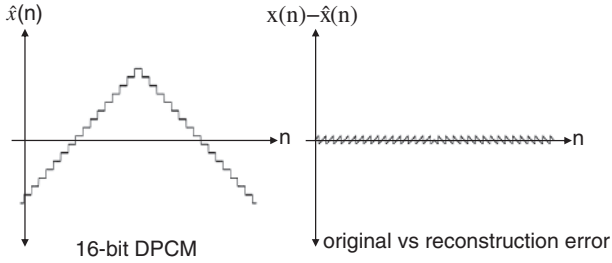


Figure 1.28 PCM Quantized Signal and Reconstruction Error.

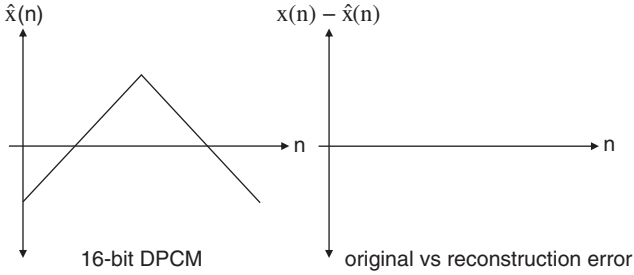


Figure 1.29 DPCM Quantized Signal and Reconstruction Error.

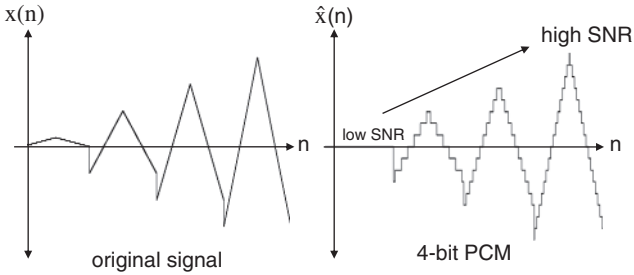


Figure 1.30 4-bit PCM.

DPCM achieves this performance gain by reducing the dynamic range of the quantized signal while preserving the number of quantization levels, thereby maintaining transmission rate efficiency.

Figure 1.30 illustrates an issue inherent to 4-bit PCM mapping. Due to the limited resolution, samples with low amplitudes undergo significant distortion during encoding, resulting in a low SNR. As the signal's amplitude increases, the relative error decreases, leading to a consequent augmentation in SNR.

Adaptive PCM (APCM) offers a method to address quantization issues that result from narrow dynamic ranges. This approach introduces adaptability by first amplifying the signal, then applying uniform quantization. This amplification occurs on a per-frame basis, and the resulting gain is encoded alongside the quantized signal samples. During decoding, the codec reverses this process by attenuating the amplified signal segments.

To elaborate further, consider a speech frame $x(n)$ comprising N 16-bit samples. The gain is determined by calculating the ratio between the dynamic range of the signal in relation to the quantizer's maximum range. This adaptive process enables to counteract quantization challenges and enhance performance. This gain, $i_g(m) \leq 1$, that is calculated for each frame m, is used to amplify the signal in accordance with $i_a(n) = \frac{x(n)}{i_g(m)}$. $N + 1$ numbers, the N signal samples $i_a(n)$, and

the gain $i_g(m)$ are transmitted to the decoder for reconstruction. The gain $i_g(m)$ is calculated as

$$i_g(m) = k_1 \sum_{n=1}^{N} x^2(n) + k_2 \tag{1.11}$$

where k_1 and k_2 are constants.

The primary challenge of APCM results from the inclusion of gain parameters in the encoding process, which slightly increases the transmission rate due to the additional data. However, speech signals typically exhibit short periods of stationarity, meaning the gain affecting one frame often closely resembles the gain affecting the previous frame. This observation leads to *Backward APCM* (BAPCM), where the gain for the current encoding frame is estimated based on the gain of the previous frame. Since both the encoder and decoder have access to the previous frame, the gain value for any given frame is simultaneously available to both endpoints. This modification allows for efficient gain adjustment while minimizing the transmission overhead associated with encoding gain parameters.

Differential pulse code modulation is a differential scheme with a fixed prediction mechanism where the predicted sample is always identical to the last encoded or decoded sample. This can be further extended by assuming a prediction that is adaptive and relies on a linear combination of past samples

$$\hat{x}_p = \sum_{i=1}^{M} a_i \hat{x}(n-i) \tag{1.12}$$

where M is the prediction order and a_i denotes prediction parameters. If adaptive quantization, like that of APCM, is combined with an adaptive prediction, a new scheme known as *Adaptive Differential Pulse Code Modulation* (ADPCM) is obtained. Both the ADPCM and DPCM encoder and decoder are very similar in that the error amplitude signal $i_a(n)$ is transmitted alongside gain parameters $i_g(m)$ computed on a frame-by-frame basis. Note that these $i_g(m)$ values extend the dynamic range of the error signal samples. Furthermore, adaptive prediction coefficients, denoted as a_i and also determined on a per-frame basis, are conveyed as $i_p(l)$. The decoder processes $i_a(n)$, $i_p(l)$, and $i_g(m)$ to yield $\hat{x}(n)$ samples. Just like any encoding and decoding mechanism within a codec, a trade-off arises between frame size–dictating how frequently these parameters are updated–and latency. Larger frames reduce overhead but concurrently increase the delay experienced between the transmitter and receiver.

Another potential enhancement to ADPCM involves integrating a backward-based estimation technique for both the gain $i_a(m)$ and the prediction coefficients $i_p(l)$. Specifically, estimation of error gain and prediction parameters is conducted using information from the preceding transmitted frame, which is accessible to both the encoder and decoder. These estimates are then applied to

the frame currently undergoing encoding or decoding. This approach gives rise to a scheme known as *Backward ADPCM* (BADPCM). Like any backward-oriented encoding mechanism, BADPCM raises a concern. Gain and prediction parameters are calculated based on previous signal samples, and these calculated values then influence the determination of current samples. This creates significant potential for error propagation. Consequently, BADPCM possesses a lower level of reliability compared to ADPCM.

1.4.1.2 LPC

Parametric codecs rely on the extraction of essential parameters from a speech sequence. These parameters are then utilized in the source decoder to synthesize the speech signal. They are also closely related to the correlation coefficients that are inherent to speech signals. Specifically, speech can be modeled as an *autoregressive* (AR) process where samples $x(n)$ are linked together as follows

$$x(n) + a_1 x(n-1) + \cdots + a_M x(n-M) = v(n) \tag{1.13}$$

and $a_1, a_2, \cdots, a_M$ are the AR parameters and $v(n)$ represents a white noise signal. Note that, therefore, the current sample value $x(n)$ is given by

$$x(n) = -a_1 x(n-1) - a_2 x(n-2) - \cdots - a_M x(n-M) + v(n) \tag{1.14}$$

as a linear combination of its own regressed values of the signal, $x(n-1), \cdots x(n-M)$, and the error term $v(n)$ that is a white noise signal plays the role of the source of air flow in the vocal tract.

Now, the following question becomes key. How can the parameters a_i that will enable the synthesis of the speech signal at the receiver be estimated? LPC analysis offers a mechanism for estimating these parameters from past speech samples. Specifically, given a frame of speech, this analysis predicts a given sample $\hat{x}(n)$ as

$$\hat{x}(n) = -\sum_{i=1}^{M} \hat{a}_i x(n-i) \tag{1.15}$$

where M is the prediction order and $\hat{a}_i$ are LP coefficients that estimate the a_i parameters. Now, the prediction error is $e(n) = x(n) - \hat{x}(n)$ so the idea is to select those $\hat{a}_i$ that minimize the error. One approach is by means of the mean-squared prediction error J as

$$J = E\{e^2(n)\} = E\left[\left(x(n) + \sum_{i=1}^{M} \hat{a}_i x(n-i)\right)^2\right] \tag{1.16}$$

where minimization results from performing partial derivatives $\frac{\partial J}{\partial a_i} = 0$ that lead to the normal equation given by

$$\sum_{i=1}^{M} \hat{a}_i R_x(i-k) = -R_x(k) \tag{1.17}$$

where $R_x(k) = \sum_{n=0}^{N} x(n)x(n-k)$ is the autocorrelation of $x(n)$ obtained over the whole N-sample frame. The normal equation can be represented as a set of M equations with M unknowns

$$\begin{aligned}
\hat{a}_1 R_s(1) + \hat{a}_2 R_s(2) \cdots \hat{a}_M R_s(M) &= -R_s(0) \\
\hat{a}_1 R_s(0) + \hat{a}_2 R_s(1) \cdots \hat{a}_M R_s(M-1) &= -R_s(1) \\
&\vdots\ \vdots \\
\hat{a}_1 R_s(-1) + \hat{a}_2 R_s(0) \cdots \hat{a}_M R_s(M-2) &= -R_s(2) \\
\hat{a}_1 R_s(1-M) + \hat{a}_2 R_s(2-M) \cdots \hat{a}_M R_s(0) &= -R_s(M)
\end{aligned} \tag{1.18}$$

that can be solved to obtain $\hat{a}_i$. One metric of interest is called *Prediction Gain* (PG) that is defined as

$$\begin{aligned}
\text{PG} &= 10 \log_{10} \left(\frac{\sigma_x^2}{\sigma_e^2} \right) \\
&= 10 \log_{10} \left(\frac{E\{x^2[n]\}}{E\{e^2[n]\}} \right)
\end{aligned} \tag{1.19}$$

and provides a ratio in dB between the energy of the input signal and the energy of the error measured by means of their corresponding variances. In general, higher PG are responsible of better signal estimation, and thus, less distortion.

To illustrate the significance of the PG, Figure 1.31 shows two speech frames, one characterized as voiced and the other as unvoiced, in terms of their temporal attributes. As expected, the voiced frame exhibits periodicity and higher energy compared to the noise-like, unvoiced frame. Furthermore, Figure 1.32 presents

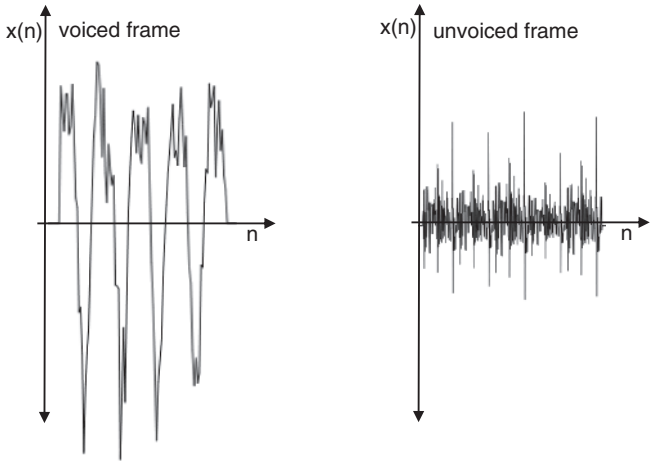


Figure 1.31 Voiced vs Unvoiced Frames.

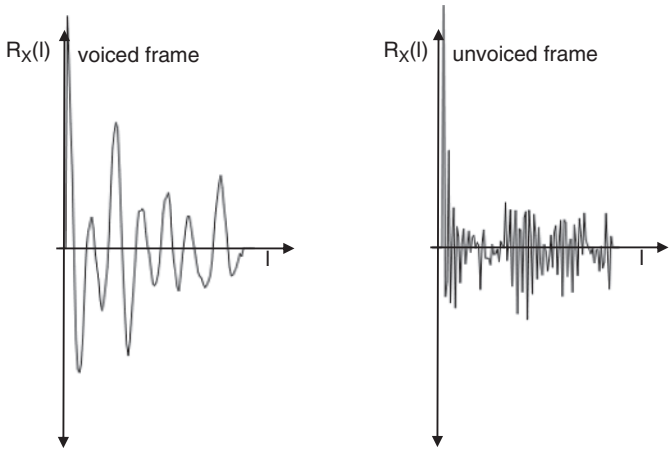


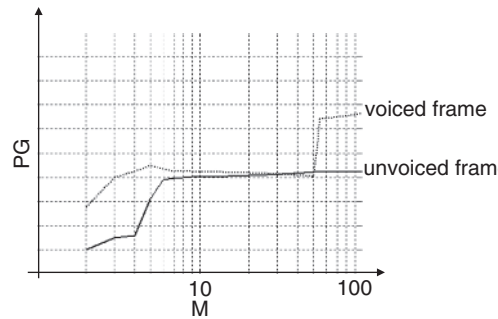
Figure 1.32 Autocorrelation.

the corresponding autocorrelation functions, denoted as $R_x(l)$ and dependent on lag, for these two speech frames. As expected from the nature of the signals themselves, voiced frames demonstrate pronounced correlation among samples, while unvoiced frames manifest randomness and minimal correlation when samples are spaced more than 5 time units apart.

When LPC is applied to the voiced and unvoiced speech frames shown in Figure 1.31 and PG is assessed in relation to the prediction order M, the results shown in Figure 1.33 are obtained. For unvoiced frames, PG increases up to $M \leq 5$ and then levels off due to the limited correlation among samples that are more distantly spaced. In opposition, for voiced frames, PG grows until $M \leq 10$, followed by a plateau until around $M \approx 50$, where it resumes growth due to the correlation among samples separated by a pitch period.

This plot confirms the observations evident in the autocorrelation plot of Figure 1.32. Unvoiced signals display minimal correlation when samples are

Figure 1.33 PG for Voiced and Unvoiced Frames.



more than 5 time units apart, while voiced signals manifest strong correlation when sample separation is within the range of 10 or greater than 50. Resolving the normal equation requires the prior calculation of autocorrelation coefficients $R_x(l)$. The timing of these calculations determines which of the two prediction mechanisms is employed (i.e., internal vs external).

In internal prediction, both the computation of autocorrelation coefficients and the LPC process itself are performed within the frame being encoded. This results in an encoded stream that includes both the LPC coefficients and the associated error signal.

Conversely, in external prediction, the encoder performs LPC on a given frame while relying on autocorrelation coefficients calculated from the previous frame. This approach assumes sufficient speech stationarity to ensure minimal variations in autocorrelation coefficients across frames. External prediction is associated with backward encoding, where past frames contribute to the computation of the current frame. Consequently, the encoded stream consists solely of the error signal, as the pre-computable LPC coefficients need not be transmitted to the decoder.

As with any backward scheme, error propagation becomes a critical concern. This issue can be mitigated by maintaining manageable frame sizes to guarantee speech stationarity while simultaneously reducing end-to-end latency.

Because PG, as shown in Figure 1.33, grows in two regions, when $M \leq 10$ and when $M \approx 50$, and remains flat for all other values of M, rather than performing an order 50 linear prediction given by

$$\hat{x}(n) = -\sum_{i=1}^{M\approx 50} \hat{a}_i x(n-i) \tag{1.20}$$

prediction can be divided into two components, short-term prediction to address the correlation within 10 samples and long-term prediction to address the correlation due to the speech pitch period. Specifically, with this modification the estimation becomes

$$\hat{x}(n) = -\sum_{i=1}^{M\approx 10} \hat{a}_i x(n-i) - \hat{a}_T x(n-T) \tag{1.21}$$

where now the prediction is order $M \approx 10$ and therefore only about 10 LP coefficients are needed besides the long-term LP coefficient $\hat{a}_T$ where $T \approx 50$. With this new scheme, a considerable improvement is done by reducing the number of encoded LP coefficients from around 50 to about 11. Depending on the speech codec under consideration, $8 \leq M \leq 12$ and $T \approx 50$ will be dynamically estimated before linear prediction is performed. In general, since the pitch period varies much faster than short-term LP coefficients, T and the long-term LP coefficient $\hat{a}_T$ must be estimated on a much faster basis.

Incorporating both short-term and long-term LPC results in a structure where frames are decomposed into approximately four or more subframes. Since short-term prediction is performed less frequently than long-term prediction, the LP coefficients a_i, with $i \leq M$, are computed on a per-frame basis, while T and a_T are computed on a per-subframe basis.

The LPC coefficients are real numbers that require quantization using a floating point scheme for encoding. Unfortunately, quantization can introduce small errors that potentially destabilize the predictor, negatively impacting the overall SNR. To address this issue, an alternative approach involves transforming LPC coefficients into a less quantization-sensitive format. Specifically, the *Line Spectral Frequency* (LSF) representation enables the conversion of LPC coefficients into LSF coefficients, which are then quantized for encoding.

1.4.1.3 LBR Codecs

While several speech codecs exist, some distinguish themselves by incorporating supplementary mechanisms that bolster overall compression efficiency. One such example is *Linear Prediction Codec 10* (LPC-10), an early and now obsolete speech codec based on LPC principles. It achieves exceptionally low transmission rates, at the expense of producing artificial-sounding speech. Operating with a frame size of 180 samples, yielding a transmission rate of 2.4 Kbps, this codec employs a rigid speech production model that categorizes frames as either voiced or unvoiced.

Like most codecs, both the encoder and decoder in LPC-10 employ preemphasis and deemphasis filters to manage high-frequency components. The preemphasis filter amplifies these components, recognizing their greater importance in speech signal intelligibility compared to lower frequencies. This results in a net improvement of the SNR. After quantization, which introduces noise into the signal, the deemphasis filter selectively attenuates the noise while preserving the amplified speech signal.

The speech production model, shown in Figure 1.34, employs three parameters, energy, zero crossing, and PG–to ascertain the voiced or unvoiced nature of a given speech segment. In the case of voiced speech, energy tends to be concentrated at lower frequencies, the signal exhibits periodic zero crossings attributed to the speech pitch, and the PG value is considerably higher compared to that of unvoiced speech.

It is important to note that this codec's binary choice between voiced and unvoiced speech segments results in either white noise or a series of pulses as input, a significant departure from the AR production model. This simplification contributes to the characteristic distortion and artificial quality often perceived in speech synthesized using LPC-10.

As previously noted, ADPCM codecs leverage LPC analysis to calculate coefficients for each frame. These coefficients, along with gain values and the resulting

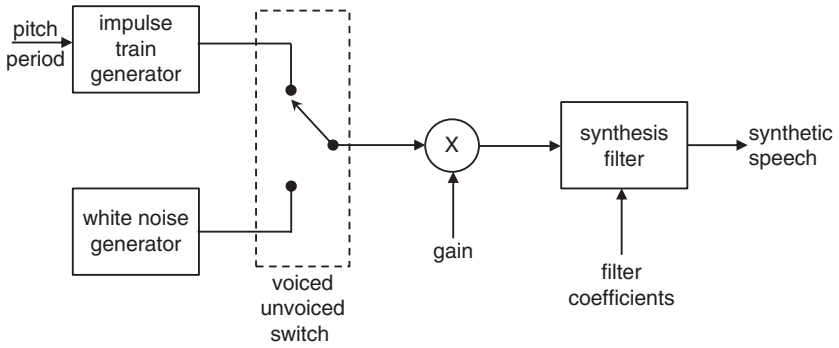


Figure 1.34 LPC-10 Speech Production Model.

error signal, are quantized to form the encoded data stream. Significantly, the error signal's samples demonstrate a relatively confined dynamic range, enabling the controlled removal of certain samples without causing perceptible distortion that would compromise audio quality.

Achieving enhanced compression becomes possible by employing a more efficient error signal encoding approach. *Code-Excited LP* (CELP), as illustrated in Figure 1.35, operates by transmitting an index that references a specific error signal or excitation within a predefined codebook or dictionary. Specifically, the CELP encoder analyzes the error signal and searches its codebook for an excitation that closely resembles it. The index associated with the chosen excitation, along with a scale factor or gain, are then encoded as components of the output stream. Conversely, the CELP decoder utilizes the index to retrieve the corresponding excitation from the codebook and subsequently rescales it using the transmitted gain value. Once the error signal is reconstructed, speech synthesis proceeds in a manner analogous to other ADPCM codecs.

More important, the synthetic speech generation process involves encoding both long-term LPC parameters, indispensable for pitch synthesis, and short-term LPC coefficients, utilized for formant synthesis. A very important requirement is ensuring absolute codebook consistency between the encoder and decoder,

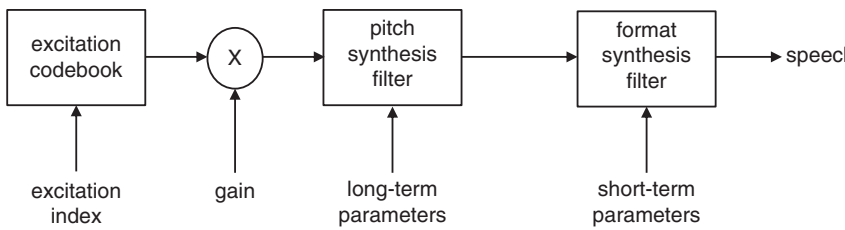


Figure 1.35 CELP.

even if these codebooks undergo dynamic construction and synchronization throughout speech processing.

The question is, how does a CELP encoder determine the index of the excitation that most closely aligns with the error signal? It accomplishes this through a technique called *analysis-by-synthesis*. This process involves systematically comparing the input error signal with a range of scaled excitations, ultimately selecting the one that yields the minimal energy of discrepancy.

For further optimization, CELP has integrated several enhancements, including techniques such as *Vector Sum Excited LP* (VSELP), which constructs codewords through the mixing of a predetermined set of basis vectors. Another CELP variant that has widespread adoption due to its simplified codebook structure is *Algebraic CELP* (ACELP). Rather than relying on extensive codebooks, ACELP employs binary addition and shifting operations to build excitation codewords. As an alternative to reduce codebook complexity, *Mixed CELP* (MCELP) mixes aperiodic, randomly perturbed pulses with noise sequences to generate excitation vectors.

In contrast, *Low delay CELP* (LDCELP) relies on small frames and subframes to facilitate external prediction. This strategy enables LPC coefficients specific to one frame to be calculated based on statistics from the preceding frame. As a result, the necessity for transmitting LPC coefficients, which can be computed concurrently by both the encoder and decoder, is eliminated. Additionally, as these coefficients are not transmitted, there is no requirement to support short- and long-term predictions, allowing for comprehensive LPC of order $M = 50$ to be conducted for improved compression efficiency.

The speech codecs discussed earlier exhibit a *Constant Bit Rate* (CBR), wherein a fixed transmission rate is determined by the ratio of bits within a frame to the frame's duration in seconds. However, in many scenarios, it is possible to encode the speech frame using fewer bits, making a fixed transmission rate less efficient. Addressing this challenge, *Variable Bit Rate* (VBR) codecs offer a solution by dynamically selecting the most efficient encoding scheme for each frame based on an analysis of speech and background noise.

1.4.1.4 ITU-T Recommendation G.711

The *International Telecommunications Union* (ITU) ITU-T Recommendation G.711 presents an HBR waveform codec encoder that transforms the 16-bit samples into 8-bit samples utilizing the established μ-Law and A-Law curves. This conversion process reduces each sample's size by half, resulting in an output transmission rate of 64 Kbps for the codec encoder. Note that unlike traditional PCM, the compression in ITU-T Recommendation G.711 occurs on a per-frame basis. Specifically, an input frame containing 80, 160, or 240 16-bit samples is converted into a corresponding output frame with 80, 160, or 240 8-bit samples, respectively (ITU-T 1988a).

Tables 1.2 and 1.3 show the procedure for converting 16-bit samples to 8-bit samples according to ITU-T Recommendation G.711 for both μ-Law and A-Law encoding. Consider a 16-bit sample *1000 0000 1001 0000*. When subjected to μ-Law encoding, the entry *s0000000wxyz'a* → *s000wxyz* applies due to the presence of the sequence *0000000* following the initial bit. In this instance, $s = 1$, $wxyz = 1001$, resulting in an output 8-bit value of *1000 1001*. Conversely, if A-Law is employed for the same input sample, the entry *s00000001wxyz'a* → *s000wxyz* comes into play. Here, $s = 1$ and $wxyz = 0010$, yielding an output 8-bit value of *1000 0010*.

Figures 1.36 and 1.37 show an ITU-T Recommendation G.711 μ-Law compression scheme and a linear compression scheme, respectively. These schemes involve the conversion of 16-bit samples into 8-bit samples with values spanning from $-2^7 + 1 = -127$ to $2^7 = 128$. When a sinusoidal signal with a varying amplitude A is subjected to each of these schemes, the resulting SNR plots shown

Table 1.2 A-Law.

16-bit Raw Input Sample	μ-law Code
s0000000wxyz'a	s000wxyz
s0000001wxyz'a	s001wxyz
s000001wxyz'ab	s010wxyz
s00001wxyz'abc	s011wxyz
s0001wxyz'abcd	s100wxyz
s001wxyz'abcde	s101wxyz
s01wxyz'abcdef	s110wxyz
s1wxyz'abcdefg	s111wxyz

Table 1.3 A-Law.

16-bit Raw Input Sample	A-law Code
s00000001wxyz'a	s000wxyz
s0000001wxyz'ab	s001wxyz
s000001wxyz'abc	s010wxyz
s00001wxyz'abcd	s011wxyz
s0001wxyz'abcde	s100wxyz
s001wxyz'abcdef	s101wxyz
s01wxyz'abcdefg	s110wxyz
s1wxyz'abcdefgh	s111wxyz

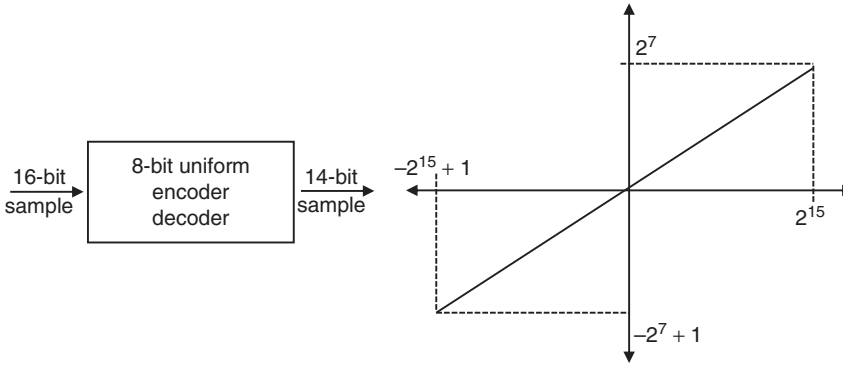


Figure 1.36 8-bit Uniform Quantization.

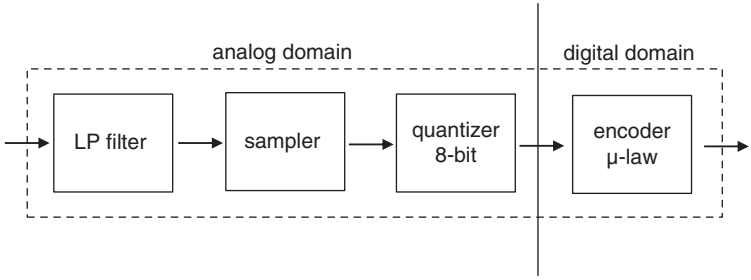
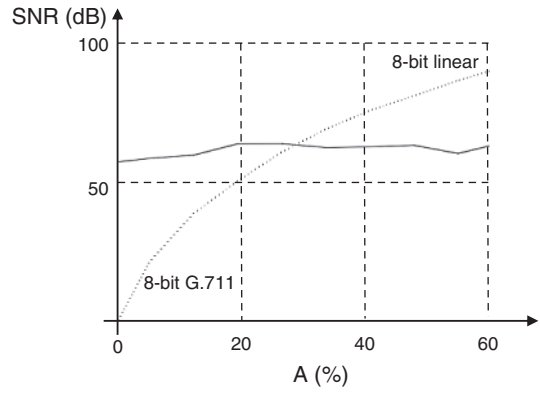


Figure 1.37 8-bit ITU-T Rec. G.711 Quantization.

Figure 1.38 Performance Comparison; Linear vs ITU-T Rec. G.711.



in Figure 1.38 are generated. Note that when the input dynamic range is about 30% of the threshold, a point is reached where lower amplitudes, commonly associated with speech signals, exhibit superior performance under the ITU-T Recommendation G.711 μ-Law compression compared to the linear compression.

The ITU-T Recommendation G.711 constitutes an NB codec, where sampling occurs at a rate of 8000 sps. With each sample being 8 bits in size, this results in a transmission rate of 64 Kbps. Because the codec processes samples in frames that are 10 milliseconds long, the frames are $\frac{64 \text{ Kbps} \times 0.01 \text{ s}}{8} = 80$ bytes long. Note that although ITU-T Recommendation G.711 is a speech codec, tones like the *Dual-Tone Multi-Frequency* (DTMF) ones used for signaling are not degraded and can be transmitted and detected.

1.4.1.5 ITU-T Recommendation G.726

Similar to how ITU-T Recommendation G.711 adopts PCM, the ITU-T Recommendation G.726 HBR speech codec employs ADPCM, offering four distinct transmission rates (16, 24, 32, and 40 Kbps) based on quantization levels. ITU-T Recommendation G.726 utilizes a second-order LPC predictor as a foundational component of its process.

The quantization levels are associated with four different numbers of bits per sample. Since ITU-T Recommendation G.726 is also an NB codec, n-bit samples result in a transmission rate of $n \times 8000$ bps. For 2-bit, 3-bit, 4-bit, and 5-bit samples, this leads to the aforementioned transmission rates. For a frame duration of 10 milliseconds, the frame sizes are respectively $\frac{16 \text{ Kbps} \times 0.01 \text{ s}}{8} = 20$ bytes, $\frac{24 \text{ Kbps} \times 0.01 \text{ s}}{8} = 30$ bytes, $\frac{32 \text{ Kbps} \times 0.01 \text{ s}}{8} = 40$ bytes and $\frac{40 \text{ Kbps} \times 0.01 \text{ s}}{8} = 50$ bytes. Because of the considerably high transmission rates supported by ITU-T Recommendation G.726, tone distortion is minimal, and DTMF tones can be transmitted and detected (ITU-T 1990).

Example 1.2 The ITU-T Recommendation G.726 codec samples speech at an NB rate. Assuming input 16-bit raw samples, what is the compression rate (the ratio between the compressed and non-compressed media rate) for 40 Kbps?

Solution:

If $F_s = 8000$ sps and $n = 16$ bits, then the raw transmission rate is $R = F_s \times n = 8000 \times 16 = 128$ Kbps. The compression rate is then $k = \frac{40}{128} = 0.3125$.

1.4.1.6 ITU-T Recommendation G.723.1

The ITU-T Recommendation G.723.1 introduces an MBR codec that relies on ACELP to support two transmission rates of 5.3 and 6.3 Kbps. Because of the low transmission rates, it is only suitable for speech, and signaling DTMF tones are distorted such that they cannot be detected. As such, this codec requires out-of-band mechanisms to support the transmission of DTMF tones.

The ITU-T Recommendation G.723.1 codec is also an NB codec with a frame duration that is 30 milliseconds long. At 8000 sps, each frame includes 8000 sps $\times$ 0.03 s = 240 samples. The frame size for 5.3 Kbps is $\frac{5300 \text{ bps} \times 0.03 \text{ s}}{8} \approx 20$ bytes

long. Similarly, the frame size for 6.3 Kbps is $\frac{6300 \text{ bps} \times 0.03 \text{ s}}{8} \approx 24$ bytes long. The transmission rates are further improved by relying on the silence compression speech scheme specified in Annex A of the ITU-T Recommendation G.723.1. This mechanism enables the support of *Discontinued Transmissions* (DTX) through *Comfort Noise Generation* (CNG) that results from *Silence Insertion Descriptor* (SID) frames (ITU-T 1996a).

1.4.1.7 ITU-T Recommendation G.729 Annexes A, D, and E

The ITU-T Recommendation G.729 Annex A delineates a well-known MBR ACELP codec, which sustains a transmission rate of 8 Kbps. Like other MBR codecs, identifying DTMF tones is unfeasible due to the distortion engendered by compression. Being an NB codec and employing a 10-millisecond frame duration results in a frame size of 10 bytes, calculated as $\frac{8000 \text{ bps} \times 0.01 \text{ s}}{8}$.

Annex A of the ITU-T Recommendation G.729 is commonly coupled with Annex B to facilitate DTX and CNG, executed through the transmission of SID frames. In addition, ITU-T Recommendation G.729 Annexes D and E respectively lower and increment the transmission rate to 6.4 Kbps and 11.8 Kbps for 8-byte and 15-byte payload sizes (ITU-T 1996b).

1.4.1.8 GSM 6.10

An important example of an early MBR codec is the GSM 6.10 codec, which introduces a technique whereby the error signal is subjected to analysis and downsampling to reduce the overall transmission rate. This method, referred to as multipulse excitation, involves encoding a limited number of non-zero samples by indicating both their positions and values.

In Figure 1.39, a modification of multipulse excitation called *Regular Pulse Excitation* (RPE) is shown. This method involves downsampling the error signal into multiple sub-sets, each defined by distinct initial offsets. The set with the greatest energy content is chosen, and its sample values, along with an identifier for the selected set, are quantized and encoded to replace the original error signal.

Table 1.4 shows the bit allocation for GSM 6.10, where short- and long-term predictions are performed over 20-millisecond frames. The GSM 6.10 codec therefore specifies a transmission rate given by $\frac{260 \text{ bits}}{0.02 \text{ s}} = 13$ Kbps. Note that this MBR codec cannot be used with DTMF tones as they cannot be detected due to the compression distortion (ETSI 1992).

1.4.1.9 AMR

The *Adaptive Multi-Rate* (AMR) codec is a multipulse excitation CELP NB codec with the capability to accommodate eight distinct transmission rates, facilitating both LBR and MBR speech encoding. The primary concept underlying this codec is its ability to dynamically select the optimal transmission rate for ensuring the

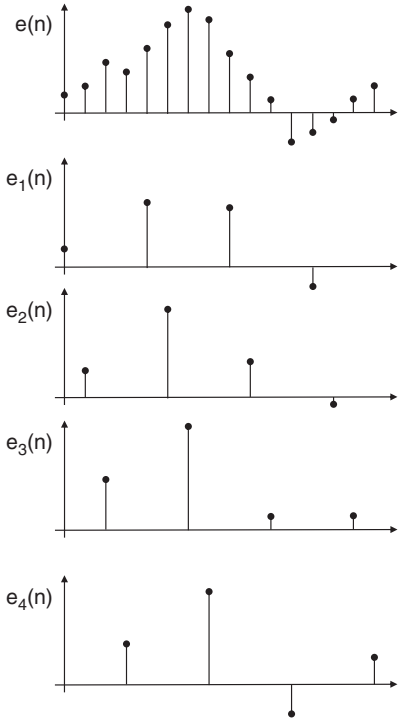


Figure 1.39 RPE.

Table 1.4 GSM 6.10 Bit Allocation

Parameter	Parameters per Frame	Resolution	Total Bits per Frame
LPC	8	6,6,5,5,4,4,3,3	36
Pitch Period	4	7	28
Long-Term Gain	4	2	8
Position	4	2	8
Peak Magnitude	4	6	24
Sample Amplitude	4-13	3	156
Total			260

fulfillment of QoS goals. These eight rates include 4.75 Kbps, 5.15 Kbps, 5.90 Kbps, 6.70 Kbps, 7.40 Kbps, 7.95 Kbps, 10.20 Kbps, and 12.20 Kbps.

For frame durations of 20 milliseconds, the corresponding frame sizes are approximately 12 bytes for 4.75 Kbps, 13 bytes for 5.15 Kbps, 15 bytes for 5.90 Kbps, 17 bytes for 6.70 Kbps, 19 bytes for 7.40 Kbps, 20 bytes for 7.95 Kbps, 26 bytes for

10.20 Kbps, and 31 bytes for 12.20 Kbps (calculated as $\frac{r \text{ bps} \times 0.02 \text{ s}}{8}$, where r is the corresponding AMR rate). The AMR codec, however, introduces signal distortions that hinder its ability to detect DTMF tones. Consequently, external methods are required to transmit such tones using out-of-band mechanisms. AMR supports DTX and also provides an additional transmission rate of 1.8 Kbps, specifically designed for transmitting SID frames. AMR is key to transmission of speech and audio in the context of the 4G technologies described in Section 4.3 (3GPP 2008a).

1.4.1.10 ITU-T Recommendation G.728

The ITU-T Recommendation G.728 describes an HBR CELP codec that supports a transmission rate of 16 Kbps. Although it is an LPC-based codec, its rate is large enough that this NB codec does not distort the media stream and can, therefore, be used to detect DTMF tones. Each ITU-T Recommendation G.728 frame is 20 milliseconds long and carries $\frac{16000 \text{ bps} \times 0.02 \text{ s}}{8} = 40$ bytes (ITU-T 1992).

1.4.1.11 ITU-T Recommendation G.722

The ITU-T Recommendation G.722 introduces an HBR *wideband* (WB) codec associated with a WB sampling rate of 16000 sps. The codec relies on an ADPCM mechanism that operates at 48, 56, and 64 Kbps. As an HBR codec, DTMF tones exhibit little distortion and they can be detected when carried in ITU-T Recommendation G.722 streams. For 20-millisecond frames, each frame carries $\frac{48000 \text{ bps} \times 0.02 \text{ s}}{8} = 120$ bytes, $\frac{56000 \text{ bps} \times 0.02 \text{ s}}{8} = 140$ bytes, and $\frac{64000 \text{ bps} \times 0.02 \text{ s}}{8} = 160$ bytes for 48, 56, and 64 Kbps (ITU-T 1988b).

1.4.1.12 iLBC

The *Internet Low Bitrate Codec* (iLBC) is an MBR codec that relies on LPC techniques to accomplish compression. It is specified in the IEFF RFC 3951. It is an NB codec that supports two frame durations: 20 and 30 milliseconds. As an MBR codec, it introduces distortion that prevents DTMF tones from being detected. For 20-millisecond duration frames, the data transmission rate is 15.2 Kbps, resulting in a frame size of 38 bytes calculated as $\frac{15200 \text{ bps} \times 0.02 \text{ s}}{8}$. Similarly, when considering frames lasting 30 milliseconds, the transmission rate becomes 13.334 Kbps, and the corresponding frame size is approximately 50 bytes, derived from $\frac{13334 \text{ bps} \times 0.03 \text{ s}}{8}$ (Hagen et al. 2004).

1.4.1.13 Speex

Speex represents a CELP-based codec that accommodates a broad spectrum of sampling and transmission rates, supporting LBR, MBR, and HBR configurations. The supported rates span from 2 Kbps to 44 Kbps. In configurations geared toward higher sampling and transmission rates, the codec is generally able to detect DTMF tones. Note that Speex extends its support to NB sampling at 8000 sps, WB

sampling at 16000 sps, *super wideband* (SWB) sampling at 32000 sps, and *Fullband* (FB) sampling at 48000 sps.

The frame duration associated with Speex is consistently set at 30 milliseconds. Consequently, the frame size ranges from around 8 bytes, calculated as $\frac{2000\ \text{bps} \times 0.03\ \text{s}}{8}$, to 165 bytes, derived from $\frac{44000\ \text{bps} \times 0.03\ \text{s}}{8}$ (Valin 2016).

1.4.1.14 EVRC and EVRC-B

The *Enhanced Variable Rate Codec* (EVRC) stands as a CELP NB speech codec, accommodating three distinct transmission rates: MBR 8.55 Kbps, LBR 4 Kbps, and 800 bps. This last rate, the lowest, serves the purpose of facilitating DTX by allowing the conveyance of SID frames. When considering a frame duration of 20 milliseconds, these rates translate to frame sizes of approximately 22 bytes (for $\frac{8550\ \text{bps} \times 0.02\ \text{s}}{8}$), 15 bytes (for $\frac{4000\ \text{bps} \times 0.03\ \text{s}}{8}$), and 2 bytes (for $\frac{800\ \text{bps} \times 0.02\ \text{s}}{8}$), respectively. Similar to AMR, the aim is for these rates to be dynamically chosen to fulfill QoS objectives. It is important to note that this codec introduces a level of distortion that hinders its ability to detect DTMF tones. EVRC-B is an enhancement to EVRC that relies on an additional 2 Kbps rate that supports 5-byte-long frames (3GPP2 2004).

1.4.1.15 GSM-EFR

GSM Enhanced Full Rate (GSM-EFR) defines an ACELP MBR NB speech codec, enabling a transmission rate of 12.2 Kbps. In frames lasting 20 milliseconds, each frame's size is approximately 31 bytes, calculated as $\frac{12200\ \text{bps} \times 0.02\ \text{s}}{8}$. Due to the distortion introduced by GSM-EFR at this transmission rate, detecting DTMF signals becomes challenging, requiring the use of out-of-band mechanisms for their successful transmission (ETSI 1999).

1.4.1.16 LPC-10

The LPC-10 codec, designated as *Federal Information Processing Standard* (FIPS) publication 137, represents an earlier LBR codec that has been deprecated but it remains important due to its historical relevance. Operating with a frame duration of 22.5 milliseconds, and a transmission rate of 2.4 Kbps, it results in a frame size of approximately 7 bytes ($\frac{2400\ \text{bps} \times 0.0225\ \text{s}}{8}$). Due to its nature as an NB LBR codec, the distortion introduced by LPC-10 prevents the accurate detection of DTMF tones. Other details of LPC-10 are presented in Section 1.4.1.3 (FIPS 1984).

1.4.1.17 AMR-WB/ITU-T Recommendation G.722.2

The speech codec known as *Adaptive Multi-Rate Wideband* (AMR-WB), which is standardized as ITU-T Recommendation G.722.2, is built upon the ACELP technique. This codec is designed to facilitate high-quality speech communication

and offers a range of nine transmission rates: 6.6 Kbps, 8.85 Kbps, 12.65 Kbps, 14.25 Kbps, 15.85 Kbps, 18.25 Kbps, 19.85 Kbps, 23.05 Kbps, and 23.85 Kbps.

When considering a frame duration of 20 milliseconds, the corresponding frame sizes are approximately 17 bytes, 23 bytes, 32 bytes, 36 bytes, 40 bytes, 46 bytes, 50 bytes, 58 bytes, and 60 bytes, respectively. These sizes are calculated using the formula $\frac{r \text{ bps} \times 0.02 \text{ s}}{8}$, where r represents the given AMR-WB rate. These characteristics categorize the codec as both MBR and HBR.

Like AMR, AMR-WB operates adaptively, dynamically selecting the optimal transmission rate based on specified QoS objectives. It should be noted, however, that the codec introduces a level of distortion that precludes the detection of DTMF tones. The codec's functionality includes support for DTX modes, accomplished through the transmission of SID frames. AMR-WB is key to transmission of speech and audio in the context of the 5G technologies described in Section 4.3 (3GPP 2008b).

1.4.1.18 RTAudio

Real Time Audio (RTAudio) is a CELP-based speech codec that supports an NB 8.8 Kbps MBR rate and WB 18 Kbps as well as 24 Kbps HBR rates. With a frame duration set at 20 milliseconds, the corresponding sizes for these frames are calculated as follows: 22 bytes for 8.8 Kbps, 45 bytes for 18 Kbps, and 60 bytes for 24 Kbps (as $\frac{r \text{ bps} \times 0.02 \text{ s}}{8}$ where r is the corresponding transmission rate). Note that, all of these bit rates introduce a level of distortion that renders proper detection of DTMF tones unfeasible within the signal, requiring the out-of-band transmission of such tones.

1.4.1.19 EVS

The *Enhanced Voice Services* (EVS) codec is a speech codec based on MBR and HBR ACELP techniques. It introduces a primary mode that offers support for various sampling rates, including NB, WB, SWB, and FB. When using NB sampling, there are seven potential transmission rates: 5.9 Kbps, 7.2 Kbps, 8 Kbps, 9.6 Kbps, 13.2 Kbps, 16.4 Kbps, and 24.4 Kbps. WB sampling allows for 12 rates: 5.9 Kbps, 7.2 Kbps, 8 Kbps, 9.6 Kbps, 13.2 Kbps, 16.4 Kbps, 24.4 Kbps, 32 Kbps, 48 Kbps, 64 Kbps, 96 Kbps, and 128 Kbps. In the case of SWB sampling, there are nine feasible rates: 9.6 Kbps, 13.2 Kbps, 16.4 Kbps, 24.4 Kbps, 32 Kbps, 48 Kbps, 64 Kbps, 96 Kbps, and 128 Kbps. Finally, for FB sampling, seven rates are available: 16.4 Kbps, 24.4 Kbps, 32 Kbps, 48 Kbps, 64 Kbps, 96 Kbps, and 128 Kbps. In addition, under WB sampling, the *AMR-WB InterOp* (AMR-WB IO) mode can accommodate all AMR-WB rates detailed in Section 1.4.1.17.

For frames lasting 20 milliseconds, the frame sizes are determined by the formula $\frac{r \text{ bps} \times 0.02 \text{ s}}{8}$, where r denotes the given rate. The resulting frame sizes are as follows: 15 bytes for 5.9 Kbps, 19 bytes for 7.2 Kbps, 20 bytes for 8 Kbps,

24 bytes for 9.6 Kbps, 33 bytes for 13.2 Kbps, 41 bytes for 16.4 Kbps, 61 bytes for 24.4 Kbps, 80 bytes for 32 Kbps, 120 bytes for 48 Kbps, 160 bytes for 64 Kbps, 240 bytes for 96 Kbps, and 320 bytes for 128 Kbps. Moreover, EVS supports DTX and CNG through the transmission of SID frames. While HBR transmission rates allow for DTMF tone detection, it is advisable to use external mechanisms for transmitting these tones to ensure reliable detection. As AMR-WB, EVS is key to transmission of speech and audio in the context of the 5G technologies described in Section 4.3 (3GPP 2018).

1.4.1.20 Lyra

Lyra, a speech codec developed by Google, employs *Machine Learning* (ML) techniques for feature extraction from the input signal to generate an encoded stream. Lyra is compatible with NB, WB, SWB, and FB sampling rates, producing LBR and MBR transmission rates of 3.2 Kbps, 6 Kbps, and 9.2 Kbps. The codec handles DTX and CNG through the use of SID frames, while DTMF tone detection requires out-of-band mechanisms. In the case of 20-millisecond frames, the frame sizes are 8 bytes, 15 bytes, and 23 bytes for rates of 3.2 Kbps, 6 Kbps, and 9.2 Kbps, respectively (calculated as $\frac{r \text{ bps} \times 0.02 \text{ s}}{8}$, where r represents the specified rate). The latest version of Lyra, V2, is optimized to improve audio quality and latency by relying on a *Residual Vector Quantizer* (RVQ) (Google 2023).

1.4.2 Audio Coding

Speech is a distinct form of audio signal, known for predominantly containing its spectral components at frequencies below 4 kHz. It can be represented through an AR process. In contrast, more intricate audio signals, such as music and tones, possess a more consistent spectral distribution that stretches up to 20 kHz. Due to their intricacy, these signals cannot be efficiently compressed using LPC techniques. As a result, audio encoders are capable of handling speech signals, while speech encoders are not suitable for processing general audio signals.

In contrast to speech, audio undergoes compression through methods involving spectral transformations like the *Modified Discrete Cosine Transform* (MDCT), which are similar to the techniques employed in video compression (in Section 1.4.3). Once transformed into the frequency domain, redundancy within the spectral samples can be eliminated by capitalizing on the human auditory system's perceptual limitations. Note that the phenomenon of audio masking influences the threshold of ear perception, shown in Figure 1.40, and dictates the extent to which data can be discarded without compromising quality. The perceptual threshold graph illustrates that the highest sensitivity of the human ear is within the sub-4 kHz range. For instance, a tone A remains audible as long as it lies above this threshold.

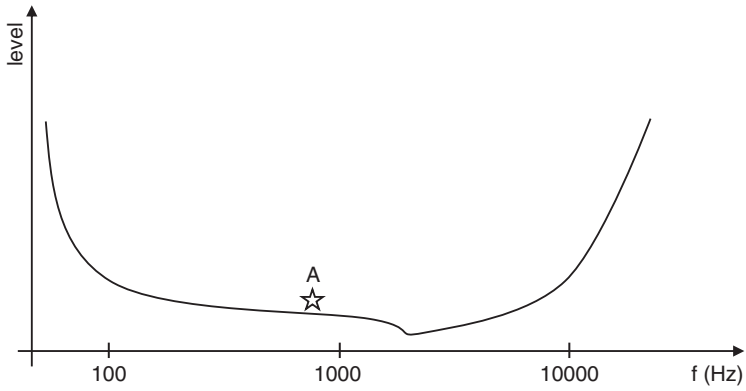


Figure 1.40 Ear Perception Threshold with Audible Tone A.

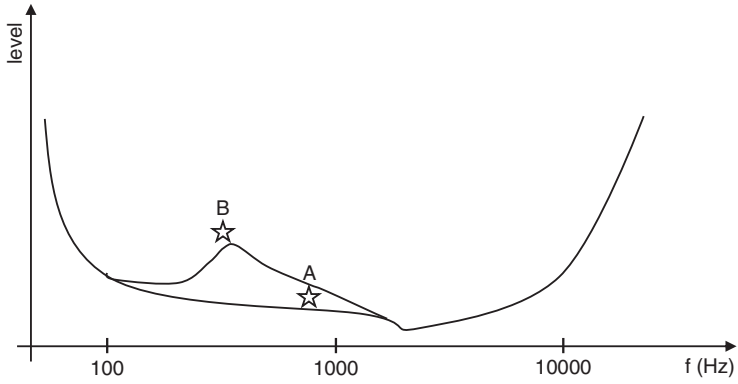


Figure 1.41 Ear Perception Threshold (Tone A is masked by tone B).

When two tones, such as tone A and tone B illustrated in Figure 1.41, have frequencies that are in close proximity, the louder tone conceals the presence of the other through a phenomenon referred to as frequency masking. Similarly, in the time domain, if a faint signal follows a powerful one or a strong signal is succeeded by a feeble one, the dominant signal eventually eclipses the other due to temporal masking (Figure 1.42).

The psychoacoustic model derived from the perception threshold serves as a cornerstone for audio encoders. This model, in turn, plays a pivotal role in shaping a perceptual encoder, which encodes incoming audio based on the sensitivity of the human ear across various spectral regions. The process, as shown in Figure 1.43, involves two sequential operations: (1) the incoming signal is passed through a bank of 32 subband filters, referred to as a bank filter and (2) the resulting 32 waves are individually subjected to quantization using quantizers that follow the

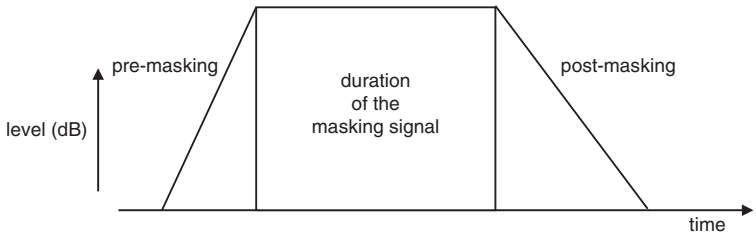


Figure 1.42 Temporal Masking.

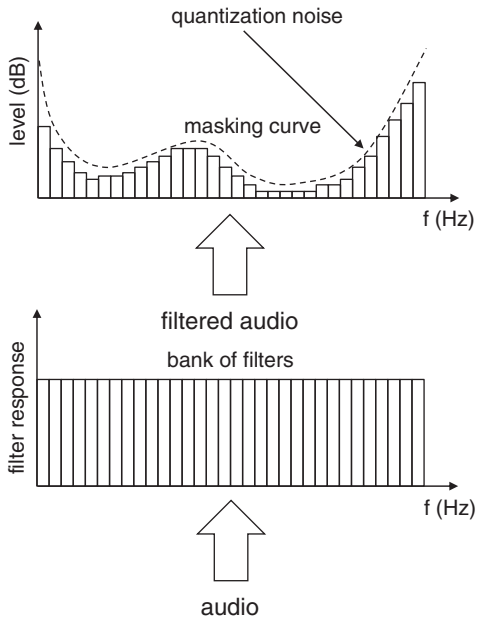


Figure 1.43 Perceptual Audio Coding.

perception threshold specific to each filter's spectral location. The underlying concept hinges on the notion that higher perception thresholds allow for greater permissible quantization errors. This implies that subband samples originating from regions with larger thresholds can be encoded with fewer bits than those from areas with smaller thresholds, thereby reducing the overall transmission rate.

A perceptual encoder can be used for stereo audio where the compression of audio originating from distinct left and right channels is performed. To achieve this, each channel undergoes the previously mentioned filter bank process. By leveraging the psychoacoustic model's assessment of masking effects, quantizers are precisely configured to operate at appropriate quantization levels. The quantizer outputs are then translated into symbol-encoded codewords, which are subsequently multiplexed to form the compressed audio stream.

Similarly, a perceptual decoder designed for stereo audio functions in a reverse manner. It starts by demultiplexing the compressed audio stream into distinct codewords, which are subsequently transformed into analog samples through a *Digital-to-Analog Converter* (DAC). These analog samples are then combined to reconstruct the audio streams for both the left and right channels.

Observing the impact of subband filtering on the overall transmission rate under the context of sampling, it can be seen that subband filtering does not influence the transmission rate. Under sampling without the inclusion of subband filtering, every sample is encoded using B bits, and W signifies the signal's bandwidth, leading to a transmission rate of Rate $= 2WB$. Similarly, when the same signal is subjected to sampling with the addition of 4-band filtering, each subband holds a bandwidth of $\frac{W}{4}$, yet the cumulative transmission rate remains unchanged, maintaining Rate $= 4 \cdot 2 \cdot \frac{W}{4} \cdot B = 2WB$.

As previously indicated, an aspect of audio compression pertains to the incorporation of multiple channels, resulting in distinct operational modes. These modes include stereo compression, where the left and right channels are individually encoded with varying quantization levels; joint stereo, which considers the redundancy present between the left and right channels during encoding; dual channel, involving the independent encoding of channels as uncorrelated waveforms; and mono, which entails encoding only a single channel.

1.4.2.1 AAC

The *Advanced Audio Coding* (AAC) standard establishes an audio codec relying on the MDCT method, allowing for sampling rates of up to 96000 sps and supporting 48 channels in joint stereo mode. This results in the existence of both CBR and VBR scenarios, which are linked to variable frame lengths. Among the three profiles supported by the standard, *AAC-Low Complexity* (AAC-LC) is capable of accommodating up to 16 channels (ISO/IEC 2006).

1.4.2.2 ITU-T Recommendation G.718

ITU-T Recommendation G.718 introduces a codec that adapts its behavior based on the sampling rates employed. When operating at NB sampling rates, it functions as a speech codec, utilizing CELP techniques for compression. In contrast, at WB sampling rates, it transitions to an audio codec, employing MDCT for compression. In speech coding modes, the codec achieves distortion levels that prevent the detection of DTMF tones. Additionally, it supports features such as DTX and CNG through the transmission and detection of SID frames.

This recommendation supports a range of transmission rates, including both MBR and HBR options: 8 Kbps, 12 Kbps, 16 Kbps, 24 Kbps, and 32 Kbps. Specifically, NB sampling rates are compatible with 8 Kbps and 12 Kbps, whereas WB sampling rates are compatible with all five transmission rates. For 20-millisecond

frames, these rates result in frame sizes of 20 bytes for 8 Kbps, 30 bytes for 12 Kbps, 40 bytes for 16 Kbps, 60 bytes for 24 Kbps, and 80 bytes for 32 Kbps (calculated as $\frac{r \text{ bps} \times 0.03 \text{ s}}{8}$, where r is the transmission rate under consideration) (ITU-T 2008).

1.4.2.3 ITU-T Recommendation G.722.1

The ITU-T Recommendation G.722.1 outlines an HBR WB audio codec that utilizes an MDCT mechanism to achieve data transmission rates of 24 Kbps and 32 Kbps. This transform-based codec exhibits minimal distortion when handling DTMF tones, and these tones are typically conveyed through separate out-of-band mechanisms. The frame duration is 20 milliseconds and, therefore, the frame sizes amount to 60 bytes and 80 bytes for transmission rates of 24 Kbps and 32 Kbps, respectively (calculated as $\frac{r \text{ bps} \times 0.03 \text{ s}}{8}$ where r is the transmission rate under consideration).

1.4.2.4 iSAC

The *internet Speech Audio Codec* (iSAC) stands as a creation from the developers behind the iLBC codec, known as *Global IP Solutions* (GIPS). iSAC serves as a complementary counterpart to iLBC by extending its capabilities to cover both speech and audio content. This is achieved through transform coding mechanisms similar to MDCT, where speech and audio signals undergo WB and SWB sampling at rates of 16000 sps and 32000 sps, respectively.

For WB sampling, the transmission rates span from approximately 10 Kbps to 32 Kbps, employing two different frame durations: 30 milliseconds and 60 milliseconds. As a result, frame sizes for a 30-millisecond duration range from about 38 bytes (calculated as $\frac{10000 \text{ bps} \times 0.03 \text{ s}}{8}$) to 120 bytes (calculated as $\frac{32000 \text{ bps} \times 0.03 \text{ s}}{8}$). These values double when considering 60-millisecond frames.

Likewise, for SWB sampling, the transmission rates vary between roughly 10 Kbps and 56 Kbps, utilizing a fixed frame duration of 30 milliseconds. This leads to frame sizes ranging from around 38 bytes to 210 bytes (calculated as $\frac{56000 \text{ bps} \times 0.03 \text{ s}}{8}$).

1.4.2.5 SILK

SILK, a speech-oriented audio codec, utilizes LPC speech compression techniques to accommodate NB, *Medium Band* (MB) at 12000 sps, WB and SWB sampling rates. As a result, it achieves MBR and HBR VBR transmission rates ranging from 6 Kbps to 40 Kbps. When considering a frame duration of 20 milliseconds, the frame sizes vary between 15 bytes (computed as $\frac{6000 \text{ bps} \times 0.02 \text{ s}}{8}$) and 100 bytes (calculated as $\frac{40000 \text{ bps} \times 0.02 \text{ s}}{8}$) (Vos et al. 2010).

1.4.2.6 Opus

The Opus codec extends the functionality of the SILK codec, as introduced in Section 1.4.2.5. It incorporates an MDCT mechanism known as the *Constrained*

Energy Lapped Transform (CELT) to enable audio compression. In addition to supporting NB, MB, WB, and SWB sampling rates, Opus also accommodates FB sampling rates. As a result, Opus provides VBR transmission rates ranging from an MBR of 6 Kbps to an HBR of 512 Kbps. The frame sizes vary between 15 bytes (calculated as $\frac{6000 \text{ bps} \times 0.02 \text{ s}}{8}$) and 1280 bytes (calculated as $\frac{512000 \text{ bps} \times 0.02 \text{ s}}{8}$) (Valin et al. 2012).

1.4.2.7 LC3

The *Low Complexity Communication Codec* (LC3) is an audio codec based on CELT, offering support for NB, WB, SWB, and FB, along with a 44100 sps sampling rate option. This codec achieves HBR VBR rates ranging from 16 Kbps to 320 Kbps for 10-millisecond frames. Consequently, the frame sizes fall within the range of 200 bytes (calculated as $\frac{16000 \text{ bps} \times 0.01 \text{ s}}{8}$) to 4000 bytes (computed as $\frac{320000 \text{ bps} \times 0.01 \text{ s}}{8}$) (Bluetooth 2018).

1.4.3 Video Coding

In the realm of digital technology, it can be thought that digital video emerges when an analog video input is transformed into a bitstream by a source encoder. This resulting bitstream can be duplicated, encrypted, stored, and adapted for transmission through a channel. Given that analog video's LP bandwidth is approximately 5 MHz, using a sampling rate of 10 Msps is deemed safe. However, this leads to excessively high raw video transmission rates. The initiative to establish standards for video digitization began in 1982 with the now obsolete ITU-R Recommendation BT.601. This recommendation, based on a 6-MHz video signal bandwidth and quantization of samples at 8 or 10 bits, relies on a sampling frequency of $F_s = 13.5$ Msps, generating minimum transmission rates of $R = 8F_s = 108$ Mbps.

The resultant digital video frames consist of either 525 or 625 lines, each containing a corresponding number of 720 luminance or 360 chrominance samples. These samples are arranged in an orthogonal sampling configuration. Nonetheless, directly sampling and quantizing analog to digital video is bounded by the quality of the original analog video. Moreover, digital video obtained through this process is not well-suited for editing, encryption, compression, or other subsequent forms of processing.

The widespread use of *Charged-Coupled Devices* (CCD) has resulted in the development of cameras that, unlike traditional analog video sampling, offer high-quality video at a reasonable cost. Furthermore, digital cameras produce images in digital raw video formats that are exceptionally well-suited for digital processing and transmission. Among these formats, one called 4:2:2 encodes four luminance samples, along with two red chrominance and two blue chrominance

samples. In this format, each luminance frame measures 720×480 pixels^2 (United States) or 720×576 pixels^2 (Europe and rest of the world), while each chrominance frame measures 360×480 pixels^2 (United States) or 360×576 pixels^2 (Europe). At 8-bit per pixel or sample, the United States-based transmission rate is given by

$$\begin{aligned} \text{rate} &= 720 \times 480 \times 30 \times 8 + \\ &\quad 360 \times 480 \times 30 \times 8 + \\ &\quad 360 \times 480 \times 30 \times 8 \\ &= 165.888 \text{ Mbps} \end{aligned} \tag{1.22}$$

and the European-based transmission rate is identical and equal to

$$\begin{aligned} \text{rate} &= 720 \times 576 \times 25 \times 8 + \\ &\quad 360 \times 576 \times 25 \times 8 + \\ &\quad 360 \times 576 \times 25 \times 8 \\ &= 165.888 \text{ Mbps} \end{aligned} \tag{1.23}$$

because although the United States format includes fewer lines (480 vs 576), these lines are encoded more often (30 fps vs 25 fps) leading to a ratio $\frac{576}{30} = \frac{480}{25} = 19.2$.

Generally, by reducing both vertical and horizontal resolutions of the 4:2:2 format, various raw video format derivatives can be generated. For instance, the 4:2:0 raw video format involves encoding an average red chrominance sample and an average blue chrominance sample for every four luminance samples. The effective vertical and horizontal resolution of each chrominance frame is halved. More precisely, each luminance frame measures 720×480 pixels^2 (United States) or 720×576 pixels^2 (Europe), while each chrominance frame measures 360×240 pixels^2 (United States) or 360×288 pixels^2 (Europe). At 8-bit per pixel or sample, the United States-based transmission rate is given by

$$\begin{aligned} \text{rate} &= 720 \times 480 \times 30 \times 8 + \\ &\quad 360 \times 240 \times 30 \times 8 + \\ &\quad 360 \times 240 \times 30 \times 8 \\ &= 124.416 \text{ Mbps} \end{aligned} \tag{1.24}$$

and the European-based transmission rate is identical and equal to

$$\begin{aligned} \text{rate} &= 720 \times 576 \times 25 \times 8 + \\ &\quad 360 \times 288 \times 25 \times 8 + \\ &\quad 360 \times 288 \times 25 \times 8 \\ &= 124.416 \text{ Mbps} \end{aligned} \tag{1.25}$$

where again, although the United States format includes fewer lines (480 vs 576), these lines are encoded more often (30 fps vs 25 fps) leading to the same ratio $\frac{576}{30} = \frac{480}{25} = 19.2$. 4:2:0 images are generated out of 4:2:2 ones and they serve as input to most digital video compression schemes. When 4:2:0 is spatially downsampled by half, on both vertical and horizontal dimensions, a format known as *Source Intermediate Format* (SIF) is obtained. In the context of digital video, this format provides a comparable, now obsolete, Video Home System (VHS)-like quality. At 8-bit per pixel or sample, the United States-based transmission rate is given by

$$\begin{aligned} \text{rate} &= 360 \times 240 \times 30 \times 8 + \\ &\quad 180 \times 120 \times 30 \times 8 + \\ &\quad 180 \times 120 \times 30 \times 8 \\ &= 31.104 \text{ Mbps} \end{aligned} \tag{1.26}$$

and the European-based transmission rate is identical and equal to

$$\begin{aligned} \text{rate} &= 360 \times 288 \times 25 \times 8 + \\ &\quad 180 \times 72 \times 25 \times 8 + \\ &\quad 180 \times 72 \times 25 \times 8 \\ &= 31.104 \text{ Mbps} \end{aligned} \tag{1.27}$$

where the overall rate is a quarter of that of 4:2:0. The *Common Intermediate Format* (CIF) was designed as a trade-off between United States and European based SIF formats. The frame rate was set at 30 fps and the luminance and chrominance resolutions at 360×288 pixels^2 and 180×144 pixels^2, respectively. At 8-bit per pixel or sample, the CIF transmission rate is given by

$$\begin{aligned} \text{rate} &= 360 \times 288 \times 30 \times 8 + \\ &\quad 180 \times 144 \times 30 \times 8 + \\ &\quad 180 \times 144 \times 30 \times 8 \\ &= 37.324 \text{ Mbps} \end{aligned} \tag{1.28}$$

slightly higher than the SIF transmission rate. By reducing the spatial resolution of CIF by half, *Quarter Common Intermediate Format* (QCIF) is obtained. The frame rate is also reduced to 15 fps. At 8-bit per pixel or sample, the QCIF transmission rate is given by

$$\begin{aligned} \text{rate} &= 180 \times 144 \times 15 \times 8 + \\ &\quad 90 \times 144 \times 15 \times 8 + \\ &\quad 90 \times 144 \times 15 \times 8 \\ &= 4.665 \text{ Mbps} \end{aligned} \tag{1.29}$$

much lower than the CIF transmission rate. When the frame rate is reduced to 7.5 fps, the corresponding QCIF transmission rate results

$$\begin{aligned} \text{rate} &= 180 \times 144 \times 7.5 \times 8 + \\ &\quad 90 \times 144 \times 7.5 \times 8 + \\ &\quad 90 \times 144 \times 7.5 \times 8 \\ &= 2.332 \text{ Mbps} \end{aligned} \tag{1.30}$$

which although lower, it is still very high specially when considering that it is associated with very poor quality.

For good quality, *High Definition* (HD) raw formats can be used instead. Although there are multiple HD formats, two of them are fairly common; the 720p scheme is a progressive scanning mechanism that provides a resolution of 1280×720 pixels^2 and the 1080i scheme is an interleaved scanning mechanism that provides a resolution of 1920×1080 pixels^2. All these formats do not make any special downsampling of the chrominance components as they are considered 4:4:4. Under 4:4:4 for every four luminance samples, four blue chrominance and four red chrominance samples are also encoded. Because of the huge amount of information that is transmitted, transmission rates typically range between 1 and 3 Gbps based on the resolution and frame rate under consideration. For a 1080i resolution at 60 fps with 8-bit per pixel or sample, the transmission rate is given by

$$\begin{aligned} \text{rate} &= 1920 \times 1080 \times 60 \times 8 + \\ &\quad 1920 \times 1080 \times 60 \times 8 + \\ &\quad 1920 \times 1080 \times 60 \times 8 \\ &= 3 \times 1920 \times 1080 \times 60 \times 8 \\ &= 2985.984 \text{ Mbps} \approx 3 \text{ Gbps} \end{aligned} \tag{1.31}$$

where it is clear that a compression scheme becomes vital. Other *HD Television* (HDTV) resolutions include the generic 4K resolutions that are typically associated with dimensions involving a horizontal size of 4000 pixels. Although they are standardized by multiple organization, the ITU-R Recommendation BT.2020 specifies the *Ultra HD TV1* (UHDTV1) resolution as 3840×2160. The same recommendation also specifies an 8K UHDTV2 resolution with dimension 7680×4320.

Video compression heavily depends on space-frequency transformations similar to those achieved through Fourier transforms. More specifically, a primary technique called the *Discrete Cosine Transform* (DCT) decomposes an input signal into sets of sine and cosine functions. These functions, in turn, can be entirely described using amplitude and phase coefficients that influence these harmonic components. From a processing perspective, an analog signal is first represented as a sequence of numbers, for example by means of sampling and quantization,

then when subjected to the DCT, a new sequence of numbers or coefficients is obtained as

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right) \tag{1.32}$$

where x_n is the input sequence and X_k is the output transformed sequence with $k = 0, \cdots N-1$. Because images have two dimensions, a more appropriate transformation is a 2D DCT one where the DCT is applied independently on the vertical and horizontal dimensions and an $N \times N$ input image is transformed into an $N \times N$ output image as follows

$$X_{k_1,k_2} = \sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1} x_{n_1,n_2} \cos\left(\frac{\pi}{N_1}\left(n_1+\frac{1}{2}\right)k_1\right)\cos\left(\frac{\pi}{N_2}\left(n_2+\frac{1}{2}\right)k_2\right) \tag{1.33}$$

where x_{n_1,n_2} and X_{k_1,k_2} are input and output pixels, respectively.

As shown in Figure 1.44, due to the significant computational demands of applying a 2D DCT to a large image, images are divided into multiple 8×8 blocks. This segmentation enables the execution of individual 8×8 DCT operations on each block. Within this process, each of the 64 input samples is transformed into 64 output samples that depict the significance of components within a set comprising of 8×8 basis blocks. The basis blocks situated closer to the upper-left corner of the set exhibit lower spatial frequencies compared to those positioned nearer to the lower-right corner.

As shown in Figure 1.45, when considering an 8×8 input block comprising quantized 8-bit samples, the resulting 8×8 output represents a set of coefficients that characterize the significance of basis blocks. The nature of high-resolution images tends to make these blocks exhibit minimal transitions, causing the weight of low spatial frequency components to generally outweigh that of high spatial

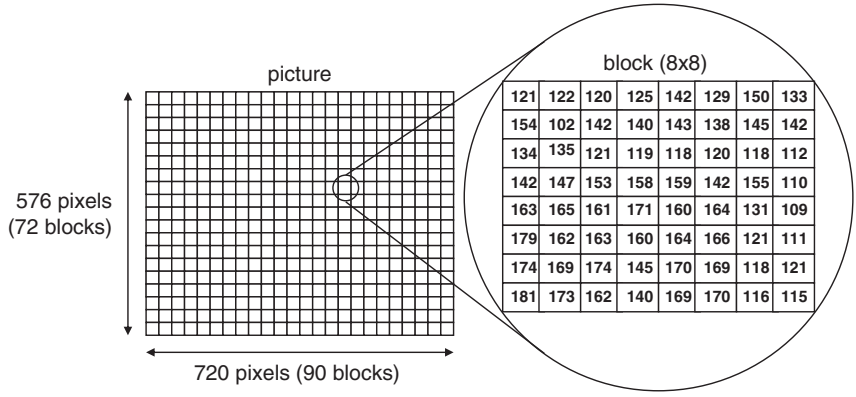


Figure 1.44 8×8 Blocks.

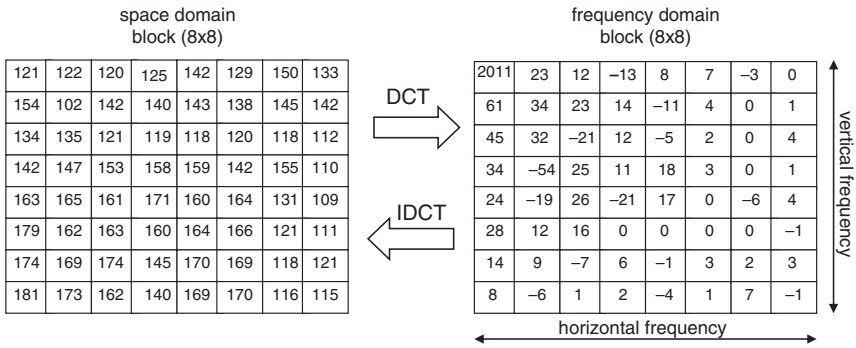


Figure 1.45 DCT of a Block.

frequency components. Consequently, the output coefficients generated by the 2D DCT experience a reduction in amplitude as frequency increases.

Additionally, it is observable that the quality of reconstruction progressively deteriorates as more components are selectively eliminated from the output sequence during the transition from the input sequence. In general, the magnitude of a coefficient correlates with the distortion introduced upon its removal. Consequently, to reduce transmission rates, it becomes feasible to encode exclusively those coefficients of sufficient size to impact human perception. Specifically, smaller coefficients that signify high spatial frequency components, which are commonly less influential, can be omitted without significantly impacting overall quality.

Hence, every coefficient can be matched against a threshold that progressively diminishes as the frequency rises. In the event that a coefficient's value falls below the threshold, it is substituted with zero. Once this thresholding procedure is enacted upon the coefficients, the resultant values are subjected to quantization using fewer levels as the vertical and horizontal frequencies ascend. This process is referred to as thresholding and quantization.

Due to the two-dimensional nature of each block, including both horizontal and vertical dimensions, and considering that transmission is a unidirectional process, a mechanism for sequential traversal and serialization of blocks becomes necessary. A favored approach is achieved through zig-zag serialization which arranges coefficients based on their spatial frequency characteristics, progressing from the lowest to the highest. It is worth noting that the very first coefficient located at the top-left corner signifies the DC component and is not tied to spatial frequency. Typically, this *Direct Current* (DC) component is encoded using DPCM, wherein the present DC sample is differentially encoded using the previous DC sample as a reference.

Due to the effects of thresholding and quantization, zig-zag serialization typically yields a unidimensional sequence that contains an extended series of zeros. This sequence can be efficiently encoded utilizing *Run Length Coding* (RLC). In the context of RLC, if a source exhibits extended sequences of identical symbols, the encoding process involves denoting the symbol itself along with the sequence's length, rather than encoding the entire sequence of symbols. It is important to note that the transmission rate can be adjusted based on thresholding and quantization parameters, potentially at the expense of image quality. This manipulation is carried out through a mechanism termed rate-control. Note that the acronym RLC is also used to describe 3GPP Radio Link Control in Chapter 4.

The *Joint Photographic Experts Group* (JPEG) is responsible for creating the widely adopted JPEG image compression standard. The input for the JPEG algorithm can take the form of either a raw image in luminance/chrominance format or an RGB format. Lossless JPEG employs LPC mechanisms, akin to those used in speech compression, to achieve compression rates of approximately 3. Conversely, lossy JPEG leverages the DCT mechanism, detailed earlier in this section, to attain compression rates of around 10.

In essence, lossy JPEG encoding entails the following steps: (1) decomposing the raw image into 8×8 luminance and chrominance blocks, (2) applying the DCT to all resulting luminance and chrominance blocks, (3) employing thresholding and quantization on the DCT coefficients, (4) traversing each block using a zig-zag pattern to maximize sample correlation, (5) employing RLC to reduce the size of the encoded sequence of symbols, and (6) employing Huffman encoding to achieve a transmission rate as close as possible to the entropy of the sequence. Conversely, the process of JPEG decoding involves a sequence of steps that reverse these six aforementioned processes.

Up until now, compression has primarily focused on individual images. When addressing video compression, a preliminary approach might involve compressing each individual frame in a manner similar to JPEG images. However, this technique proves to be inefficient as it overlooks the substantial correlation present between consecutive frames. The *Motion Pictures Experts Group* (MPEG) has introduced the MPEG standards, which serve as a comprehensive specification including various standards for audio and video multiplexing and compression. The initial MPEG standard is the MPEG-1 standard, which acts as a foundational basis for most video compression methodologies.

MPEG-1 Part 2 video compression yields a transmission rate of approximately 1 Mbps. It builds upon JPEG-based image compression but incorporates additional mechanisms to exploit the substantial correlation between successive frames, reducing the quantity of information that requires encoding. Specifically, with the utilization of prediction combined with motion compensation, a video

frame is encoded in relation to a preceding frame by encoding the discrepancy between the two frames using a motion estimator.

The motion estimation process executed at the encoder involves a trade-off between the processing speed and video quality. Faster processing speed generally results in reduced latency. The latency is also influenced by the number of video frames or pictures grouped together for motion estimation. In practical terms, when correlations between video frames are identified, these frames need to be buffered to determine their interdependencies before the encoding process can start. The cumulative latency stemming from this grouping approach is directly proportional to the number of frames within the group, with the rate of this increase being inversely related to the video frame rate.

A *Group of Pictures* (GOP) includes various frame types categorized based on how they can be predicted from preceding frames. Specifically, there are three main types: (1) Intraframe pictures (I frames) that are encoded without external reference using a mechanism similar to JPEG compression, (2) Interframe pictures (P or predicted frames), predicted from a previous I or P reference using motion estimation techniques, and (3) Bidirectional pictures (B frames), predicted as an interpolation of future and past I and P pictures.

I frames offer the highest quality but have the lowest compression rate, followed by P frames and then B frames. Typically, the encoder dynamically compresses raw video frames and determines the type of compressed frame to generate based on rate-control and quality prerequisites. More precisely, the encoder selects the GOP size and determines which types of frames to include, resulting in a specific transmission rate.

It is important to note that at the initiation of a GOP, the initial frame is always an I frame because there are no preceding pictures for reference. When accessing a video stream randomly, the process typically occurs on a GOP-to-GOP basis, ensuring complete decoding of a picture without introducing artifacts.

Shown in Figure 1.46 is an illustrative GOP, which can be characterized by two parameters, M and N. In this context, N signifies the GOP size, set at $N = 12$ for this instance, while M represents the gap between I and P frames, which is $M = 3$ in this particular scenario. Fundamentally, the compressed GOP originates from a set of 12 raw frames, transformed into a sequence comprising compressed I, B, B, P, B, B, P, B, B, P, B, B, and I pictures, where the last I picture pertains to the subsequent GOP. However, the sequence of these raw frames must follow a certain order that aligns with the prediction hierarchy. In specific terms, the raw frames require encoding or decoding in a sequence that results in I, P, B, B, P, B, B, P, B, B, I, B, and B for transmission. This rearrangement of pictures is required due to buffering and gives rise to additional latency.

The MPEG video compression standard includes hierarchies of both subsets and supersets of GOPs. More precisely, a GOP can be further divided into frames (also

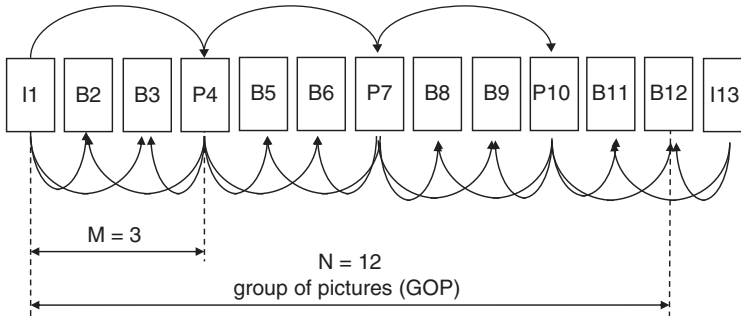


Figure 1.46 GOP Example.

known as pictures), each categorized as intra, inter, or bidirectional as previously outlined. These pictures, in turn, are divided into slices, which consist of consecutive sequences of *macroblocks* (MBs) typically organized in rows. Macroblocks generally comprise 2×2 blocks, serving as the foundational units for motion estimation. Much like the JPEG case, each block measures 8×8 pixels^2. Note that MB is also used to denominate Medium Band in Subsection 1.4.2.5.

Conversely, GOPs act as the fundamental building blocks for video sequences. All these interrelated components and their connections are represented in Figure 1.47. As a video frame is composed of numerous MBs, and considering that each picture can adopt a 4:2:0 format, it follows that each MB also conforms to the 4:2:0 format. With this arrangement, each MB includes 16×16 square luminance pixels and two sets of 8×8 square chrominance pixels.

In most instances of video compression, where MBs play a pivotal role, the picture's resolution is designed to be an integer multiple of the MB's resolution. For instance, the aforementioned CIF format relies on a resolution of 360×288, but since 360 is not an exact multiple of 16, the adapted resolution that accommodates video compression becomes 352×288.

Motion estimation involves the task of identifying identical objects in both a reference frame and the frame undergoing compression, as illustrated in Figure 1.48. Once the object is recognized, the only pertinent information to encode pertains to the object's position within the reference frame, the motion vector detailing the object's translation, and the encoding of the image portions that become exposed through this motion. As previously mentioned, object identification occurs on a per-MB basis through a search area where the reference frame is compared against the frame being considered, thereby determining the motion vector. Given that this matching process is influenced by color, it requires simultaneous execution across all chrominance and luminance components of an MB.

Since the object in the compressed frame is substantially correlated but not entirely identical to that in the reference frame, in P frames, the disparities

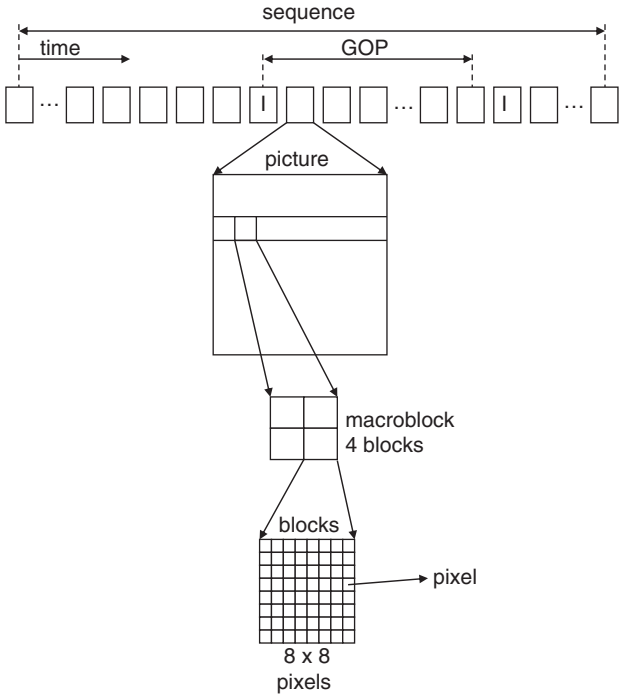


Figure 1.47 Video Frames.

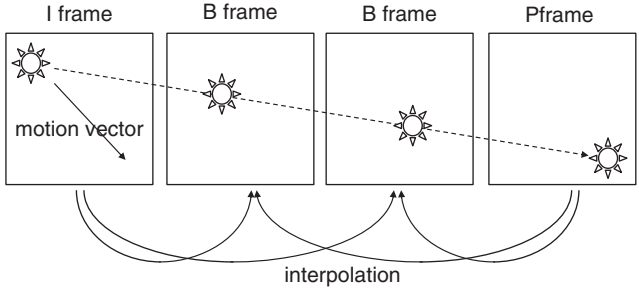


Figure 1.48 Motion Estimation.

between them must also be encoded. This is achieved through a process referred to as motion compensation. On the other hand, B frames emerge from spatial interpolation between the two reference frames, resulting in minimal information being encoded. In scenarios where the distortion resulting from B frame encoding surpasses a certain threshold, encoding is executed with P frames. Likewise, when

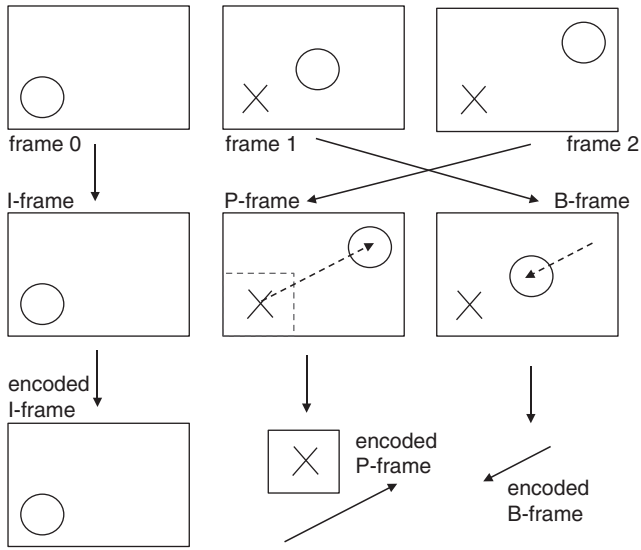


Figure 1.49 Sequence Encoding.

the distortion originating from P frame encoding exceeds a specific threshold, I frame encoding is engaged instead.

For instance, consider Figure 1.49, which illustrates a sequence featuring a circle moving along a straight path, with the intention of encoding three frames. These frames are rearranged in such a way that the initial frame is designated as an I frame, while the second and third frames are interchanged, facilitating the processing of P frames before B frames. In this arrangement, the I frame is subjected to conventional still image compression for the entire frame.

As for the P frame, it undergoes regular still image compression for the portion of the image that differs from the I frame. Additionally, a vector is incorporated to signify the direction and magnitude of the circle's movement between frame 0 and frame 2. Lastly, the B frame focuses exclusively on encoding the vector that represents the circle's movement intensity and direction between frame 2 and frame 1. Quality manipulation plays an important role in achieving rate control. Specifically, quantization parameters (such as the number of levels) and prediction schemes are dynamically chosen using a feedback mechanism to achieve a targeted transmission rate.

1.4.3.1 ITU-T Recommendation H.261

The ITU-T Recommendation H.261 specifies, a now obsolete, early VBR video codec standard that supports transmission rates between 40 Kbps and 2 Mbps to

enable both QCIF and CIF resolutions. Frame rates 7.5, 15, and 30 fps are supported (ITU-T 1993).

1.4.3.2 ITU-T Recommendation H.262 (MPEG-2 Part 2)

MPEG-2 Part 2, standardized by ITU-T Recommendation H.262, introduces a generic video coding mechanism. Within the context of MPEG-2, a set of levels is defined, linked to the resolution of the video undergoing compression. Precisely, there are four attainable levels: a low level corresponding to a resolution of CIF (360×288), a main level corresponding to 4:2:0 (720×576) resolution, a high level with 1440 resolution corresponding to standard HDTV resolution of (1440×1152), and a high level corresponding to widescreen HDTV resolution of up to (1920×1152). This leads to VBR transmission rates between 3.5 Mbps and 6 Mbps at 7.5, 15, 25, and 30 fps.

The picture types I, B, and P are present in MPEG-2. Additionally, there is a spatially scalable profile that involves transmitting a base quality picture and supplementary information layers to enhance quality. The high profile is specifically designed for HDTV encoding.

In the MPEG-2 main profile and level, interlaced 4:2:0 formatted frames with dimensions of 720×576 and captured at a frame rate of 25 fps are compressed using I, P, and B encoding (ITU-T 1995).

1.4.3.3 ITU-T Recommendation H.263

The ITU-T Recommendation H.263 enhances the previous ITU-T Recommendation H.261 and ITU-T Recommendation H.262 standards, to specify a VBR video codec that supports 7.5, 15, and 30 fps. Transmission rates between 40 Kbps and 2 Mbps are typical. Resolutions include: 128×96 (Sub Quarter Common Intermediate Format SQCIF). QCIF, CIF, 704×576 (4CIF), and 1408×1152 (16CIF).

In turn, ITU-T Recommendation H.263v2 (or H.263+) improves the original ITU-T Recommendation H.263 by adding additional annexes to the specification. Among many things, H.263+ supports custom size picture encoding (ITU-T 2005).

Example 1.3 Consider a 128 Kbps CIF ITU-T Recommendation H.263 video stream with 4:2:0 frame encoding. If the frame rate is 15, what is the overall compression rate (the ratio between the compressed and non-compressed media rate)?

Solution:

The raw video rate is given by $R = 15 \times 360 \times 288 \times 8 + 2 \times 180 \times 144 \times 8 = 12.85$ Kbps. The compression rate is then given by $k \approx \frac{12.8}{128} = 0.1$.

1.4.3.4 ITU-T Recommendation H.264 (MPEG-4 Part 10)

MPEG-2 Part 2 has been enhanced as MPEG-4 Part 10, also known as ITU Recommendation H.264 or *Advanced Video Coding* (AVC). This new standard enhances

compression efficiency by at least 50%. The primary innovation of H.264/AVC lies in its dual-layer structure: (1) the *Video Coding Layer* (VCL) integrates interframe compression, capitalizing on temporal statistical dependencies between consecutive frames, and intraframe compression, exploiting spatial dependencies within a given frame, and (2) the *Network Abstraction Layer* (NAL) packages VCL data into units, ensuring reliable transportation through data networks.

The efficiency of ITU-T Recommendation H.264 is attributed to three factors: (1) incorporation of more sophisticated prediction modes that can vary within a frame and reference a greater number of successive frames, (2) adoption of an alternative transform known as the integer transform, which is faster and employs block sizes of either 4×4 or 8×8 pixels, and (3) utilization of adaptive Huffman and adaptive algebraic coding as entropy encoding mechanisms.

The standard defines levels that support the following resolutions: QCIF, CIF, 720×576, 1280×720, 2048×1024, 3672×1536, 4096×2304, and 8192×4320. Frame rates range from 15 fps to 120 fps leading to transmission rates between 64 Kbps and 800 Mbps (ITU-T 2009).

1.4.3.5 ITU-T Recommendation H.265

The ITU-T Recommendation H.265, also known as *High Efficiency Video Coding* (HEVC), is a video compression standard that extends the ITU-T Recommendation H.264. As with this latter standard, the ITU-T Recommendation H.265 supports the following resolutions: SQCIF, QCIF, CIF, 640×360, 720×576, 960×540, 1280×720, 1920×1080, 2048×1080, 3840×2160, 4096×2160 (4K), 7680×4320 (8K), and 8192×4320 (also 8K). Transmission rates of 30 Kbps to 800 Mbps for frame rates between 15 fps and 300 fps are possible (ITU-T 2019).

1.4.3.6 ITU-T Recommendation H.266 (MPEG-I Part 3)

MPEG-I Part 3, standardized as ITU-T Recommendation H.266 and also known as *Versatile Video Coding* (VVC), supports the same resolutions and frame rates as ITU-T Recommendation H.265, including 15360×8640 (which falls under the umbrella of 16K resolutions), but improves performance and provides the same quality at lower transmission rates (ITU-T 2023).

1.4.3.7 AU1

An issue associated with compression methods based on MPEG standards pertains to their regulation by *Intellectual Property* rights. Note that a significant number of individual encoding techniques within these standards are patented, requiring specific licenses for their utilization. Addressing this concern, VP8 emerges as an open-source video codec that is devoid of patent restrictions. The primary objective of VP8 is to deliver comparable quality to ITU-T Recommendation H.264 while avoiding the need for licensing. VP8 excels in its incorporation of efficient

intra- and interprediction mechanisms, employs 4 × 4 block-oriented DCT compression for all luminance and chrominance error signals, and distinguishes itself by utilizing three reference frames for inter prediction, rather than the conventional two.

VP8 has been extended by VP9 to provide quality comparable to that of ITU-T Recommendation H.265. *AOMedia Video 1* (AU1), in turn, was developed as the successor of VP9 as an open, royalty-free video coding format that supports the following resolutions: 426 × 240, 640 × 360, 854 × 480, 1280 × 720, 1920 × 1080, 3840 × 2160, and 7680 × 4320. For frame rates between 30 fps and 120 fps, these resolutions lead to transmission rates between 1.5 Mbps and 800 Mbps for Open Media (2021).

1.4.3.8 Theora

Theora is an early open source license-free video codec derived from VP3, an ancestor of AU1. It provides similar capabilities to those of ITU-T Recommendation H.262 supporting QCIF, CIF, 4CIF, and HD resolutions of 1440 × 1152 and 1920 × 1152 (xiph.org 2006).

1.4.3.9 MPEG-5 Part 2

MPEG created the MPEG-5 Part 2 standard, also known as *Low Complexity Enhancement Video Coding* (LCEVC), to define a novel enhancement layer that builds on top of the information transmitted over a traditional base codec (i.e., AU1, ITU-T Recommendation H.264, ITU-T Recommendation H.265, ITU-T Recommendation H.266, etc.). This enhancement layer improves QoS by reducing the artifacts introduced by the base codec without significantly increasing the transmission rate (Choi et al. 2020).

1.5 Quality Scores

Due to its significant influence on user experience, the assessment of speech quality is of utmost importance. Specifically, the deterioration in quality caused by factors like packet loss, latency, and jitter is about the need for mechanisms that can evaluate this decline effectively. One such metrics is the SNR, that relies on calculating the ratio between the power associated with the source and error signals, and it is given by

$$\begin{aligned} \text{SNR} &= 10\log_{10}\left(\frac{\sum_n x_n^2}{\sum_n e_n^2}\right) \\ &= 10\log_{10}\left(\frac{\sum_n x_n^2}{\sum_n (x_n - \hat{x}_n)^2}\right) \end{aligned} \tag{1.34}$$

where x_n, e_n, and $\hat{x}_n$ are the input, the error, and the reconstructed samples of the speech under analysis, respectively. Because this method relies on signal characteristics, its precision is confined to waveform codecs exclusively. Additionally, an issue with SNR is that, on average, signals with high amplitudes tend to overshadow those with low amplitudes. This leads to coefficients that are disproportionately larger than anticipated, thus yielding an inaccurate evaluation of the analyzed speech. An alternative approach involves dividing speech streams into frames, allowing for individual SNR calculations within each segment. Then a new metric, *Segmental SNR* (SSNR) is calculated as

$$\text{SSNR} = \frac{1}{N}\sum_{m=1}^{N}\text{SNR}_m \tag{1.35}$$

as an average over all N segments. The idea is that low segmental SNR cannot be masked by high segmental SNR regions anymore.

Considering the nature of SNR and SSNR metrics, their application in evaluating parametric speech codecs yields low outcomes. This is primarily due to these codecs' tendency to distort waveforms while maintaining a level of quality. In such scenarios, an alternative metric becomes necessary, that is, one metric that involves subjective ratings gathered by averaging individuals' opinions during comparisons between source and reconstructed speech waveforms.

From a human standpoint, the perception of speech quality includes various aspects: intelligibility (how easily the speech can be understood), naturalness (the presence of undesirable elements like echoes, or noise and clicking in the reconstructed speech), and recognizability (whether the original speaker can be identified).

Moreover, in the context of parametric codecs, speech quality becomes influenced by several factors. These factors include the speaker's characteristics, the language spoken, signal intensity, background noise levels, the occurrence of channel errors, the presence of non-speech signals such as music, and tandem coding, which refers to processes like decoding, encoding, or transcoding speech across intermediary nodes.

The evaluation of speech can be conducted through various tests, depending on the chosen criteria. One such method is the *Absolute Category Rating* (ACR), where a group of listeners assesses a speech sequence and assigns rankings based on its absolute quality, ranging from excellent (5) to bad (1). The collective average of these ratings yields the *Mean Opinion Score* (MOS).

Conversely, an alternate approach known as *Degradation Category Rating* (DCR) focuses on evaluating the extent of degradation rather than quality. In this process, listeners gauge the degree of degradation experienced in the analyzed speech sequence relative to the original reference. This assessment yields five possible scores: not degraded (5), minimally degraded (4), slightly degraded (3), degraded

(2), and highly degraded (1). Once again, the average scores across all listeners are calculated, resulting in a *Degradation MOS* (DMOS).

Due to its presentation of the reference followed by the degraded speech sequence, the DCR method can introduce bias into the DMOS. An alternative assessment approach, called *Comparison Category Rating* (CCR), involves presenting listeners with both the reference and the degraded speech sequences in a randomized manner. This results in potential scores such as: the first sequence being much better than the second (3), the first sequence being better than the second (2), the first sequence being slightly better than the second (1), both sequences being identical (0), the second sequence being slightly better than the first (−1), the second sequence being better than the first (−2), and the second sequence being much better than the first (−3). As previously, these scores are averaged across all listeners.

As a general trend, increased involvement of listeners tends to yield better and more precise outcomes. Nevertheless, these evaluations demand significant time and resources, making it advantageous to devise an algorithmic solution. This algorithm would take input in the form of speech reference and degraded sequences, generating an objective output score closely aligned with the subjective score that a CCR test would yield.

For this purpose, a multitude of speech reference sequences must be utilized, each subjected to various degradation sequences. This enables the algorithm to generate objective scores, which can then be compared to corresponding subjective scores associated with the sequences.

One of the initial efforts to establish standardized speech quality assessment was undertaken through ITU-T Recommendation P.861. This recommendation outlines the algorithm called *Perceptual Speech Quality Measure* (PSQM). The block diagram framework of PSQM is illustrated in Figure 1.50. In this process, the input speech, acting as the reference, undergoes consecutive encoding and decoding

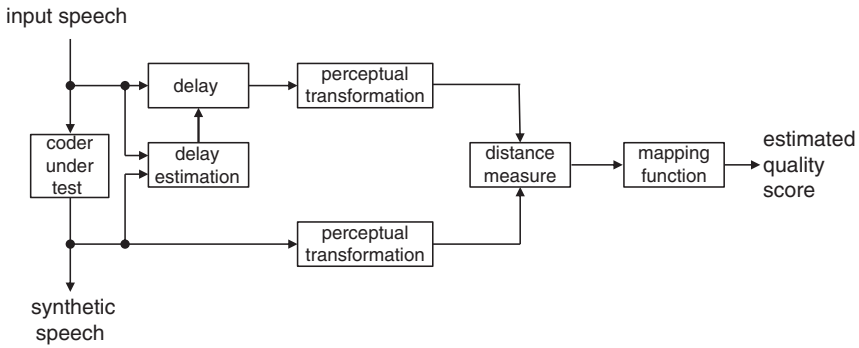


Figure 1.50 Perceptual Speech Quality Measure.

steps to generate synthetic speech. The reference signal is then temporally shifted to account for the latency introduced by codec processing. Utilizing perceptual transformations, parameters are extracted from both the reference and synthetic speech. These parameter differences are then translated into MOS speech quality scores, ranging from 1 (bad) to 5 (excellent).

1.5.1 Network Impairments

Communication networks are made of hosts that can communicate with each other. This communication is performed with the assistance of routers that are in charge of propagating information throughout the network. Both hosts and routers form communication systems with transmitters and receivers connected to channels. Each router supports multiple hosts that, in turn, are connected to sources and sinks. The network provides an alternative to dedicated links between hosts as routers allow for resource sharing in an efficient manner.

Communication networks can be either circuit switching or packet switching-based. Figure 1.51 shows a now obsolete traditional *Plain Old Telephone Service* (POTS) example of circuit switching, where there is a fixed full-duplex path between two hosts. Each telephone acting as a host is connected to a *Digital Loop Carrier* (DLC) that acting as a router allows the speech to enter the *Public Switching Telephone Network* (PSTN) where *Central Offices* (COs) or telephone exchanges

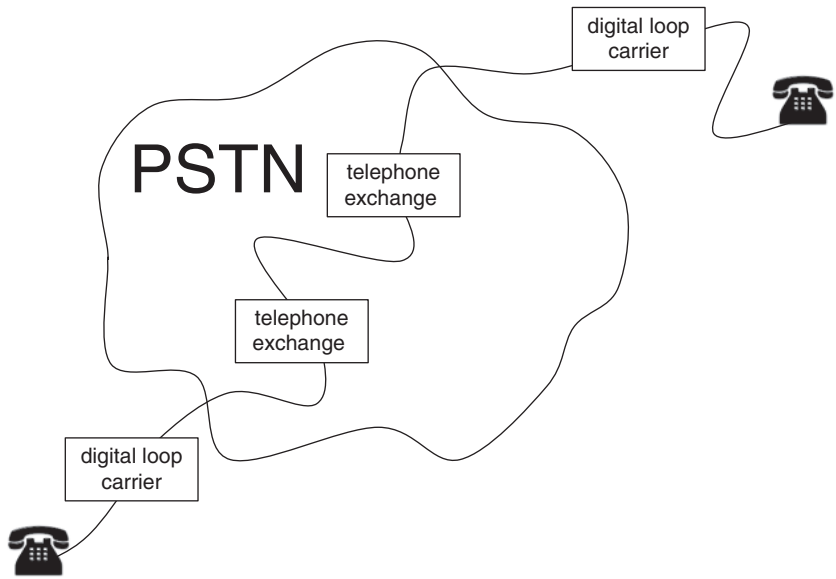


Figure 1.51 Circuit Switching.

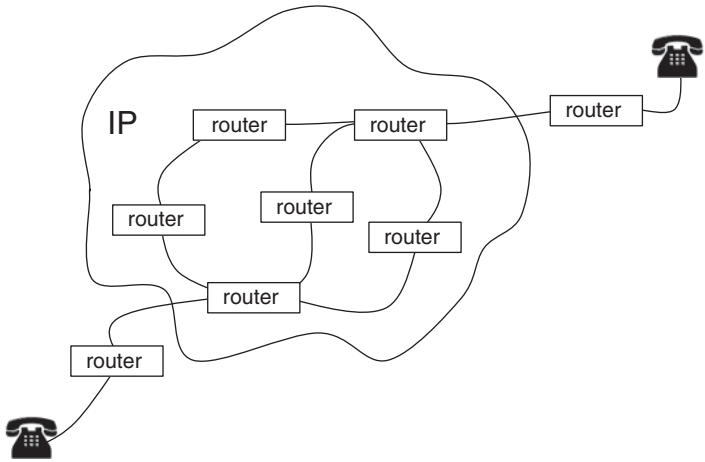


Figure 1.52 Packet Switching.

acting also as routers propagate the signals. A path consists of a common channel that traverses time slots of multiple links that once set up remain uninterrupted. When the path is established, bandwidth and other resources are assigned to the source and destination hosts until it is terminated. Because setup is fast and the path is dedicated, circuit switching is very efficient for transmission of speech.

Figure 1.52 shows an example of telephony based on packet switching to support traditional *Voice over Internet Protocol* (VoIP). Again, phones act as hosts which are now connected to IP routers, switches, and other components like *Session Border Controllers* (SBCs) acting as routers. Speech is chunked, or packetized, and then transmitted throughout the network as packets that follow a specific path from source to sink. Moreover, for each packet, the path is dynamically selected, resulting from network router conditions, congestion, and other factors that determine which set of links are used. Because the path changes on a per-packet basis, source-to-sink delay and other impairments like loss affect each packet differently, imposing quality restrictions on speech. For the same conditions, circuit switching provides much better quality than packet switching, at the cost of less efficient use of available resources. The effect of loss and latency typical of packet switching can be seen in Figure 1.53. Since Host 1 (H1) has some application data that is to be sent to Host 2 (H2), it first divides the data into four packets and adds routing headers (R) to each of them. Router 1 (R1) receives the packets and due to load balancing it forks two of them to Router 2 (R2) and the remaining two to Router 3 (R3). R3 is heavy loaded so it drops packet number 2 and only forwards packet number 4 to Router 4 (R4). Due to a slower link and higher latency, packets 1 and 3 take longer to arrive to R4 than packet 4 such that H2 receives an out of sequence partial number of packets.

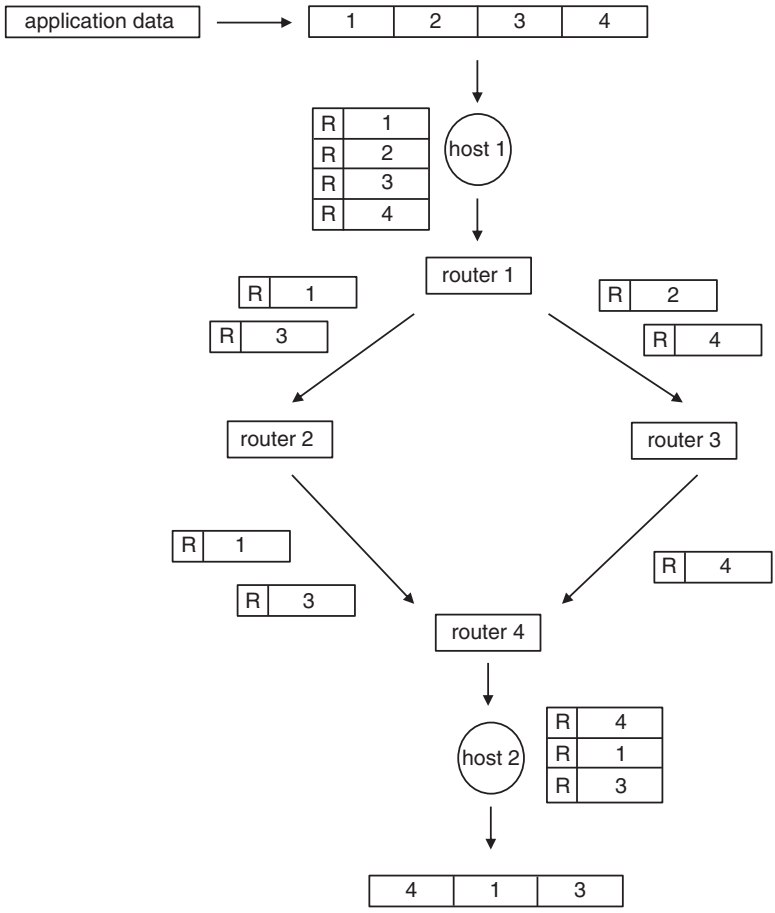


Figure 1.53 Packet Loss and Latency.

Data network communications are based on a layered architecture where each layer is a black box, such that its input is the output of the immediately lower layer, and similarly, its output becomes the input of the immediately higher layer. This exchange of information between layers is performed through each layer's interfaces, where input and output facilities exist. Each layer provides a specific function defined by the *Open Systems Interconnection* (OSI) model. The model is organized into seven layers: (1) Physical Layer handles the transmission and reception of data over the channel, including the conversion of information between the digital and analog domains, involving, therefore, physical phenomena, (2) Data Link Control provides error control mechanisms for reliable transmission of information over the channel, (3) Network is in charge of ensuring packets are delivered

from source to sink, (4) Transport provides delivery of messages between source and sink, including support for multiplexing, also known as muxing, (5) Session manages multiple sessions between endpoints, (6) Presentation formats information for further processing, and (7) Application is the layer closest to the end user and is typically involved with services and also conversion of information between the digital and analog domain.

Each layer defines a protocol that governs the conversion of information between the input and output at any given layer interface. Both Layer 1 and Layer 7 have one interface in the analog domain and another interface in the digital domain. In the analog domain, the interface deals with physical phenomena like electrical signals in the channel or acoustic signals coming from a speaker, as in RTC. Layers 2 through 6 are purely digital layers with both interfaces in the digital domain. Since humans and the real world are analog by definition, a conversion between the digital and analog domains is always critical.

The biggest advantage of a layered architecture is the compartmentalization of a communication system, such that when a specific functionality needs to be upgraded, no other components need to be changed. This leads to technologies like *Network Function Virtualization* (NFV) that provide, among other things, shorter deployment times.

The Internet is an example of a data network where different protocols are used to provide services independently at different levels, such that upgrading one layer does not affect another. For scalability reasons, the Internet uses the concept of subnetworks or subnets. Hosts belonging to the same subnet exchange information directly, and routers provide access to hosts in other subnets. Internet does not follow the Open Systems Interconnection (OSI) but instead a simplified 5-layer architecture known as the Internet Engineering Task Force (IETF) architecture. Details of this architecture are presented in Chapter 2.

Although the Internet is an example of packet switching, where each packet takes a different path based on network conditions, protocols at the transport layer like *Transport Control Protocol* (TCP) can provide a circuit switching-like behavior by associating each connection with a stream of information through virtual circuits. Simpler mechanisms like *User Datagram Protocol* (UDP), which preserve the connectionless nature of packet switching, transmit information on a per-packet or datagram basis. Figure 1.54 shows both of these transport layer protocols in the context of the network layer *Internet Protocol* (IP).

Voice over Internet Protocol applications rely on UDP and TCP transport protocols. Unfortunately, the latter involves retransmissions to guarantee data integrity, introducing delays that significantly affect end-user perception. On the other hand, UDP is affected by packet loss as well as delays inherent to packet switching, which degrade quality and affect the overall QoS. Acceptable voice

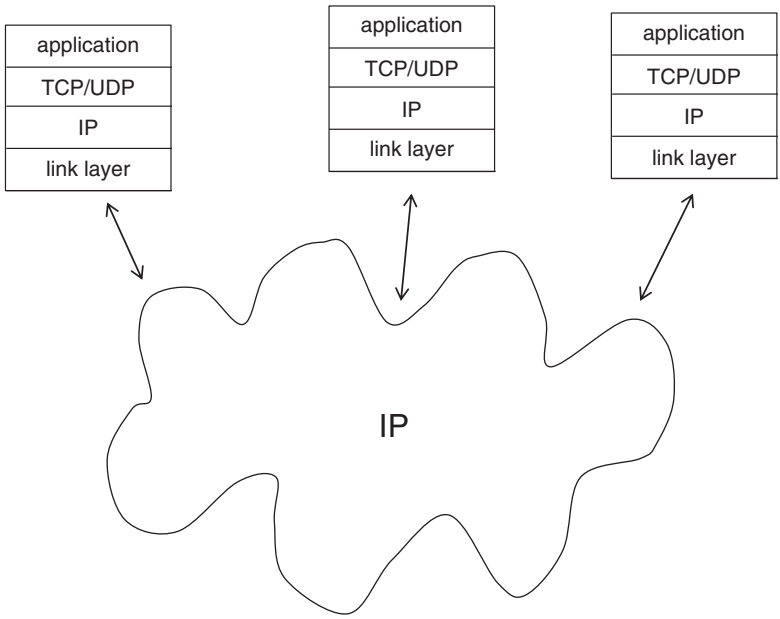


Figure 1.54 IP Protocol Layers.

quality is typically obtained when packet loss is below 3% and latency or delay is 150 milliseconds.

Figure 1.55 shows a time diagram of the effect of network delay in VoIP. If the transmitter sends a sequence of packets (1 through 8), it arrives a bit later due to network delay. Although the interpacket transmission time is constant, the interarrival time at the receiver is not always constant, leading to network jitter. To minimize the effect of both delay and jitter, the application layer uses a jitter or playout buffer to ensure that the media sequence can be played smoothly without choppiness or dead air.

1.5.2 PESQ

A refinement to PSQM emerged as *Perceptual Evaluation of Speech Quality* (PESQ), which incorporates more precise perceptual transformations, leading to enhanced accuracy in generating quality MOS. PESQ has been standardized as ITU Recommendation P.862.1 for NB speech and as ITU Recommendation P.862.2 for WB speech. Speech sampled at higher rates like SWB and FB must be downsampled to WB before PESQ can be applied. Note that because downsampling introduces distortion that lowers the scores, scores for SWB and FB sampled signals cannot be accurately obtained.

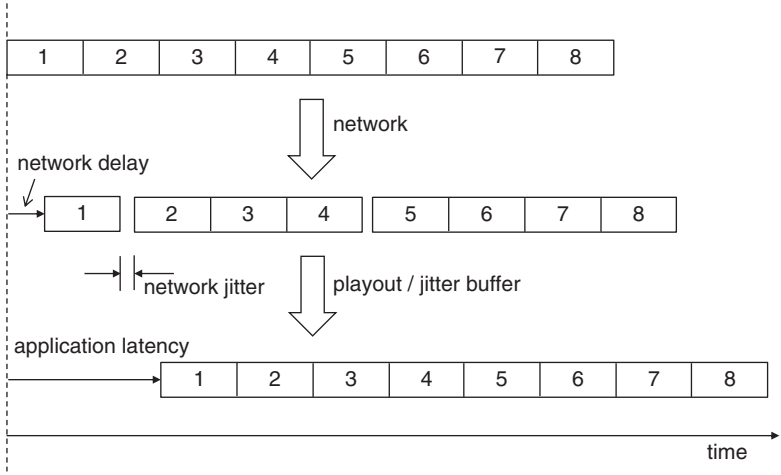


Figure 1.55 Network Delay and Jitter.

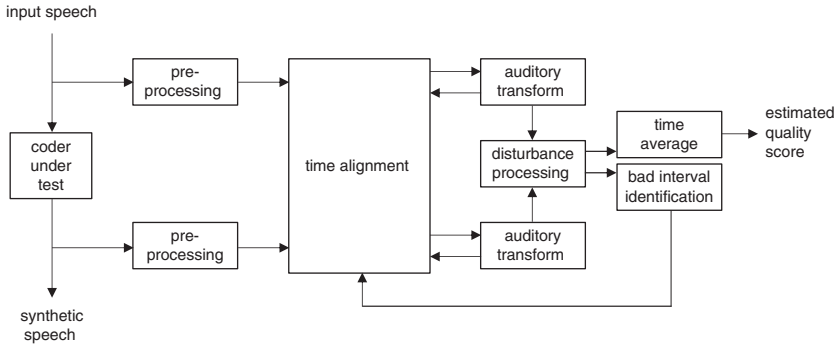


Figure 1.56 PESQ Block Diagram.

Figure 1.56 shows the block diagram of the PESQ scheme. The reference and the signal under test are respectively subject to pre-processing and then time-aligned. These signals are then transformed and auditory-processed. Time averages that lead to the actual PESQ scores are then obtained. PESQ raw scores fall in the −0.5 to 4.5 interval, but they are mapped into *MOS-Listening Quality Objective* (MOS-LQO) values in the 1 to 5 range associated with MOS described in Section 1.5. Specifically, the MOS scale assigns 1 to bad quality, 2 to poor quality, 3 to fair quality, 4 to good quality, and 5 to excellent quality. Typically, different codecs with different grades and levels of compression exhibit different MOS-LQO scores (Herrero and Cadirola 2014; ITU-T Recommendation P.862 2001).

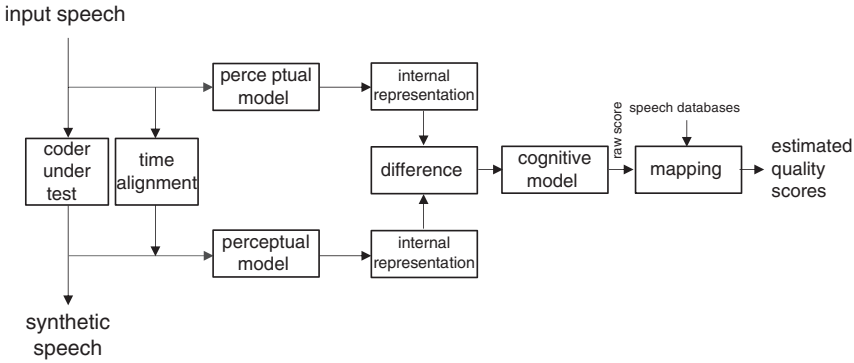


Figure 1.57 POLQA Block Diagram.

1.5.3 POLQA

In 2011, an even more accurate algorithm known as *Perceptual Objective Listening Quality Assessment* (POLQA) was standardized by the ITU through Recommendation P.863 as an evolution of PESQ. Apart from offering improved quality estimation for certain scenarios, POLQA also supports a variety of sampling rates beyond the conventional NB and WB options, including SWB and FB.

Figure 1.57 shows the different blocks involved in POLQA estimations. The reference and the signal under test are time-aligned and subjected to a perceptual model that internally represents critical parameters. The difference between these internal representations is fed to a cognitive model, which produces a POLQA raw score that is then mapped into MOS-LQO scores (ITU-T Recommendation P.863 2018).

1.6 Summary

Media generation and compression are key to RTC. Without understanding how these mechanisms work together, it is not possible to understand RTC in general. The three main types of media (speech, audio, and video) have different characteristics that make them unique from a signal processing perspective and, therefore, compression since compression depends on the nature of the signals. Speech signals are typically compressed through techniques of linear prediction and machine learning in the time domain, while audio and video respectively rely on unidimensional and bidimensional frequency domain analysis. These compression mechanisms lead to the development and use of media codecs.

Speech codecs range from low compression rate ones like the ITU-T Recommendation G.711 to high compression and performance ones like the EVS. Audio

codecs, on the other hand, involve lower compression rates to preserve quality. Examples of audio codecs include AAC and Opus. Video codecs have been standardized by the ITU with different compression rates depending on quality goals and resolution. Based on the compression ratio associated with each codec, better or worse media quality can be obtained. Objective quality metrics like PESQ and POLQA can be used to obtain scores that represent the quality of the decoded speech when compared to the originally generated one.

1.7 Homework Problems and Questions

Problem 1.1 The ITU-T Recommendation G.711 codec samples speech at an NB rate. How many 10-millisecond frames are packetized if 240 samples per packet are encapsulated?

Problem 1.2 If the EVS codec is configured at 7.2 Kbps and frames at 20 milliseconds long, how big are 3-frame EVS-encapsulated packets?

Problem 1.3 For an audio signal with bandwidth of 13 kHz, what is the best sampling rate (NB, WB, SWB, or FB)? Why?

Problem 1.4 Under ITU-T Recommendation G.726 with a transmission rate of 16 Kbps, if the sampling rate is NB, how many bits per sample is the raw speech encoded as?

Problem 1.5 Consider a 512 Kbps 1920 × 1080 ITU-T Recommendation H.265 video stream with 4:2:2 frame encoding. If the frame rate is 30, what is the overall compression rate (the ratio between the compressed and non-compressed media rate)?

Problem 1.6 What is the relationship between the raw video transmission rate at 4:2:2 encoding and the width, height, and frame rate of the video stream?

Problem 1.7 What is the trade-off between transmission rate, sampling rate, and quantization level in waveform codecs ?

Problem 1.8 Consider an LPC-based codec, if 16 4-bit coefficients are packed together to encode 10-millisecond frames, what is the transmission rate of the codec?

Problem 1.9 In an NB scenario, if ITU-T Recommendation G.711 exhibits better PESQ/POLQA scores than EVS, why is EVS the standard codec for 5G communications?

Problem 1.10 What are the advantages and disadvantages of circuit switching compared to packet switching in communication networks?

Problem 1.11 What is the purpose of PESQ in voice quality assessment? How does PESQ work?

Problem 1.12 Why do waveform codecs achieve higher PESQ/POLQA scores than LPC ones?

Bibliography

3GPP. (2008a). TS 26.071: Mandatory speech codec speech processing functions; AMR speech codec; general description. Technical Report TS 26.071, 3rd Generation Partnership Project.

3GPP. (2008b). TS 26.190: Speech codec speech processing functions; adaptive multirate – wideband (AMR-WB) speech codec; transcoding functions. Technical Report TS 26.190, 3rd Generation Partnership Project.

3GPP. (2018). TS 126.442: Mandatory speech codec speech processing functions; AMR speech codec; general description. Technical Report TS 126.442, 3rd Generation Partnership Project.

3GPP2. (2004). C.s0014-a: Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital systems. Technical Report C.S0014-A, 3rd Generation Partnership Project 2.

Alliance for Open Media. (2021). Av1 features. Available from: https://aomedia.org/av1-features/. (Accessed March 15th, 2024)

Benott, H. (2008). *Digital television*. 3rd edition. Boston: Focal Press, pp. vii–viii. ISBN 978-0-240-52081-0. DOI: https://doi.org/10.1016/B978-0-240-52081-0.50002-X. Available from: https://www.sciencedirect.com/science/article/pii/B978024052081050002X. (Accessed March 15th, 2024)

Bluetooth. (2018). An introduction to Bluetooth LE Audio. Available from: https://www.bluetooth.com/blog/a-technical-overview-of-lc3/. (Accessed March 15th, 2024)

Choi, K., Chen, J., Rusanovskyy, D., Choi, K.P., and Jang, E.S. (2020). An overview of the MPEG-5 essential video coding standard [standards in a nutshell]. *IEEE Signal Processing Magazine*, 37(3), pp. 160–167. DOI:10.1109/MSP.2020.2971765.

Chu, W.C. (2003). *Speech coding algorithms: foundation and evolution of standardized coders*. John Wiley & Sons, Ltd. ISBN 978-0-471-66885-5.

ETSI. (1992). GSM full rate speech transcoding. ETSI.

ETSI. (1999). Enhanced full rate (EFR) speech transcoding. ETSI.

FIPS. (1984). FIPS Pub 137, analog to digital conversion of voice by 2,400 bit/second linear predictive coding. FIPS.

Google. (2023). Lyra (codec). Available from: https://github.com/google/lyra. (Accessed March 15th, 2024)

Hagen, R., Andersen, S.V., Linden, J., Astrom, H., Duric, A. and Kleijn, W.B. (2004). Internet Low Bit Rate Codec (iLBC). RFC 3951, December. Available from: https://www.rfc-editor.org/info/rfc3951. (Accessed March 15th, 2024)

Haykin, S. (2009). *Communication systems*. Wiley Publishing. 5th ed. ISBN 0-471-69790-7.

Herrero, R. (2021). Fundamentals of IoT communication technologies. *Textbooks in telecommunication engineering*. Springer International Publishing. ISBN 978-3-030-70079-9. Available from: https://books.google.com/books?id=k70rzgEACAAJ. (Accessed March 15th, 2024)

Herrero, R. (2023). Practical internet of things networking. *Textbooks in telecommunication engineering*. Springer International Publishing. ISBN 978-3-031-28443-4.

Herrero, R. and Cadirola, M. (2014). Effect of FEC mechanisms in the performance of low bit rate codecs in lossy mobile environments. In *Proceedings of the Conference on Principles, Systems and Applications of IP Telecommunications, IPTComm '14*, New York, NY, USA. Association for March 2024 Computing Machinery. ISBN 978-1-450-32124-2. DOI: 10.1145/2670386.2670387. Available from: https://doi.org/10.1145/2670386.2670387. (Accessed March 15th, 2024)

ISO/IEC. (2006). ISO/IEC 13818-7: Advanced audio coding (AAC). Technical Report 13818-7.

ITU-T. (1988a). G.711: Pulse code modulation (PCM) of voice frequencies. ITU-T.

ITU-T. (1988b). G.722: 7 kHz audio-coding within 64 kbit/s. ITU-T.

ITU-T. (1990). G.726: 40, 32, 24, 16 kbit/s adaptive differential pulse code modulation (ADPCM). ITU-T.

ITU-T. (1992). G.728: Coding of speech at 16 kbit/s using low-delay code excited linear prediction. ITU-T.

ITU-T. (1993). H.261: Video codec for audiovisual services at p x 64 kbit/s. ITU-T.

ITU-T. (1995). H.262: Information technology – generic coding of moving pictures and associated audio information: Video. ITU-T.

ITU-T. (1996a). G.723.1: Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s. ITU-T.

ITU-T. (1996b). G.729: Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP). ITU-T.

ITU-T. (2005). H.263: Video coding for low bit rate communication. ITU-T.
ITU-T. (2008). G.718: Frame error robust narrow-band and wideband embedded variable bit-rate coding of speech and audio from 8-32 kbit/s. ITU-T.
ITU-T. (2009). H.264: Advanced video coding for generic audiovisual services. ITU-T.
ITU-T. (2019). H.265: High efficiency video coding. ITU-T.
ITU-T. (2023). H.266: Versatile video coding. ITU-T.
ITU-T Recommendation P.862. (2001). Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Technical Report, International Telecommunication Union, Geneva, Switzerland, November.
ITU-T Recommendation P.863. (2018). Technical Report, International Telecommunication Union, Geneva, Switzerland, September.
Jurafsky, D. and Martin, J.H. (2000). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. 1st ed. USA: Prentice Hall PTR. ISBN 0-130-95069-6.
Proakis, J.G. and Manolakis, D.K. (2006). *Digital signal processing*. 4th ed. Upper Saddle River, New Jersey: Prentice Hall. ISBN 0-131-87374-1. Available from: http://www.amazon.com/Digital-Signal-Processing-John-Proakis/dp/0131873741%3FSubscriptionId%3D192BW6DQ43CK9FN0ZGG2%26tag%3Dws%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0131873741.
Smith, J.O. (2010). *Physical audio signal processing*, Online Book. 2010 ed. W3K Publishing. ISBN 978-0-974-56072-4.
Valin, J-M. (2016). Speex: A free codec for free speech. CoRR, abs/1602.08668. Available from: http://arxiv.org/abs/1602.08668.
Valin, J-M., Vos, K. and Terriberry, T.B. (2012). Definition of the Opus 115 Audio Codec. RFC 6716, September. Available from: https://www.rfc-editor.org/info/rfc6716. (Accessed March 15th, 2024)
Vos, K., Jensen, S.S. and Soerensen, K.V. (2010). SILK Speech Codec. Internet-Draft draft-vos-silk-02, *Internet Engineering Task Force*, September. Available from: https://datatracker.ietf.org/doc/draft-vos-silk/02/. (Accessed March 15th, 2024).
xiph.org. (2006). Theora video compression. Available from: xiph.org/theora/. (Accessed March 15th, 2024)